PL-TR-95-2141

A HYPOTHESIS-TESTING APPROACH TO OUTLIER DETECTION

W.A. Woodward
H.L. Gray

Southern Methodist University
Department of Statistical Science
Dallas, TX 75275

October 1995

Final Report
September 1993 - September 1995

Approved for public release; distribution unlimited

PHILLIPS LABORATORY
Directorate of Geophysics
AIR FORCE MATERIEL COMMAND
HANSCOM AIR FORCE BASE, MA 01731-3010

SPONSORED BY
Advanced Research Projects Agency (DoD)
Nuclear Monitoring Research Office
ARPA ORDER No. A-128

MONITORED BY
Phillips Laboratory
CONTRACT No. F19628-93-C-0199

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either express or implied, of the Air Force or the U.S. Government.

This technical report has been reviewed and is approved for publication.

JAMES F. LEWKOWICZ
Contract Manager
Earth Sciences Division

JAMES F. LEWKOWICZ
Director
Earth Sciences Division

This report has been reviewed by the ESC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS).

Qualified requestors may obtain additional copies from the Defense Technical Information Center. All others should apply to the National Technical Information Service.

If your address has changed, or if you wish to be removed from the mailing list, or if the addressee is no longer employed by your organization, please notify PL/IM, 29 Randolph Road, Hanscom AFB, MA 01731-3010. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document requires that it be returned.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No: 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
October 1995

3. REPORT TYPE AND DATES COVERED
Final Report, Sept 93 - Sept 95

4. TITLE AND SUBTITLE
A Hypothesis-Testing Approach to Outlier Detection

5. FUNDING NUMBERS
F19628-93-C-0199
PE62301E
PRNM93
TAGM
WUAP

6. AUTHOR(S)
W.A. Woodward
H.L. Gray

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Southern Methodist University
Department of Statistical Science
Dallas, TX 75275

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Phillips Laboratory
29 Randolph Road
Hanscom AFB, Massachusetts 01731-3010
Contract Manager: James Lewkowicz/GPEH

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
PL-TR-95-2141

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

In this report we develop a general robust near-optimal methodology for testing for outliers to one or more populations. The methodology requires no distributional assumptions and allows imperfect data sets, i.e., data sets with missing observations. The data can be from one or several seismic stations or can be a mixture of seismic and nonseismic data and continuous and categorical data. Listings of the computer code required to implement the methodology are included.

14. SUBJECT TERMS
Likelihood ratio tests, outlier detection, bootstrapping, discriminant analysis, missing data.

15. NUMBER OF PAGES
128

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
SAR

CONTENTS

I. Summary

A. Task Objectives

Below we list the original statement of work.

1. Develop new statistical methodologies, parametric and nonparametric, which are particularly applicable to the problems of discriminant analysis, outlier detection, script matching and wave form matching in the context of monitoring nuclear proliferation.

2. Determine better methods for estimating statistical distributions which may be used for both discrimination purposes and for assessing system performance.

3. Develop a framework in which the results of monitoring and the capability of the monitoring network can be usefully and correctly stated.

4. Apply the above developed methodology to data at the ARPA Center for Seismic Studies to assess the effectiveness of the above theoretical developments.

Item 1 was accomplished by developing several near optimal tests to determine when observations should be regarded as "unusual." A paper developing a nonparametric methodology for discriminating between two groups has been accepted for publication in the journal Computational Statistics and Data Analysis. An additional paper extending our outlier detection methodology to the important missing data scenario was distributed as a technical report. Computer code to implement these results can be obtained upon request by contacting Dr. H.L. Gray, Department of Statistical Science, Southern Methodist University.

In addition, the results in outlier detection were extended to two outliers from a mixture. For example, this latter test would allow one to test for an outlier from a training set made up of mining explosions and earthquakes, rather than just one or the other. Although this latter work is not 100% complete, it is to the state of completion that it can be used in most settings. The code for this program has also been passed on to MRC and is available upon request by contacting Dr. Gray. A paper on this new outlier detection is being prepared for submission for publication.

Other methodologies were also developed. Although the theory is basically now developed, these latter methodologies are not ready for distribution.

Regarding items 2 and 3, the bootstrap methodology was introduced to effectively solve both of those problems. To satisfy item 4, the outlier method developed under this contract was applied to nuclear explosions, mining blasts, and earthquakes in diverse geological regions recorded by the ARCESS and GERESS arrays, CDSN station WMQ, and LNN stations KNB and MNV. Most such tests were run at MRC although some were also performed at SMU. At the .01 significance level, between 90-100% of the nuclear explosions and quarry blasts were detected as outliers of the earthquake groups in the various regions. Overall, 209 of 229 (91%) explosions were detected and there were only 2 false alarms out of 143 earthquakes (1.4%), not significantly higher than the targeted 1%. These results were obtained for diverse regions, for a wide range of epicentral distances and magnitudes, and for single stations and arrays. The methodology is, of course, applicable to multiple stations, as well.

The application of the outlier detection method to data from multiple stations was explored in detail, due to the concern that some data compression might be required. Various data compression methods were considered and it was ultimately decided that with proper computer code the so-called "full vector" MLE outlier method was preferable to any compression methods. This problem is discussed in detail in the paper "Outlier Tests with Multiple Stations" which is included in the appendix.

In general, we feel this work has been very successful and when the methodology we are currently developing is complete, we feel that the statistical methodology developed will be nearly optimal for automated detection of suspicious events, while at the same time furnishing the user with reliable estimates of the associated error rates.

APPENDIX

1. "A Bootstrap Generalized Likelihood Ratio Test in Discriminant Analysis," J. Baek, H.L. Gray, W.A. Woodward to appear Computational Statistics and Data Analysis.

2. "A Hypothesis-Testing Approach to Discriminant Analysis with Mixed Categorical and Continuous Variables When Data Are Missing," J.W. Miller, W.A. Woodward, H.L. Gray, M.A. Fisk, G.D. McCartor. Technical Report SMU/DS/TR-273, July 1994.

3. "Outlier Tests With Multiple Stations," H.L. Gray, W.A. Woodward, Z.T. Yucel.

4. "A New Test for Outlier Detection from a Multivariate Mixture Distribution," S. Wang, W.A. Woodward, H.L. Gray, S. Wiechecki.

A Bootstrap Generalized Likelihood Ratio Test in Discriminant Analysis

J. Baek, H. L. Gray, W. A. Woodward, J. Miller and M. Fisk

Abstract

A generalized likelihood ratio test is developed for classification in two populations when one needs to control one of the probabilities of misclassification. The proposed classification procedure is constructed by applying the parametric bootstrap to the generalized likelihood ratio. There are known methods for controlling this misclassification probability for the case where normal distributions with the same covariance matrix are assumed. Our approach, however, can be applied to not only this case but to the case of normal distributions with different covariance matrices and the case of a mixture of discrete and continuous variables.

The results given here do not depend on normality but can, in fact, be applied to any distribution for which the maximum likelihood estimates exist. We do, however, restrict our simulation of these results to the normal distribution if the variates are all continuous. Three cases are simulated: normal distributions with equal covariance matrix, normal distributions with unequal covariance matrices, and mixture of categorical and normal variables. An application to classifying seismic events is presented.

Keywords: Bootstrap, hypothesis testing, discriminant analysis, mixtures of continuous and discrete variables, mixed variables

This research was partially supported by ARPA Contracts F19628-90-C-0135, F29601-91-K-DB25, and F19628-93-C-0199.

1. Introduction

One of the primary problems associated with monitoring worldwide nuclear proliferation is the problem of distinguishing seismically between small earthquakes and explosions. Although the statistical problem appears to be one of discriminant analysis, it is actually one of testing hypotheses since the political and physical environment will usually require one of the errors to be preassigned.

Classical approaches for discriminant analysis in two populations depend on the ratio of the probabilities or probability density functions. The classification rule based on the ratio is optimal in the sense that it minimizes the total probability of misclassification (Welch 1939). Under the assumptions of normality, equal covariances, and unknown parameters for the variables, Anderson (1951) derived a classification rule based on the linear discriminant function, which is known as Anderson's W statistic, by substituting estimates for the parameters in the ratio. When the covariance matrices are not equal, replacing each parameter by its estimate gives the classical quadratic discriminant function (Seber, 1984, p297; Anderson, 1984, p235).

Among other classification rules is a hypothesis-testing approach which is derived by obtaining the generalized likelihood ratio. This rule based on the assumption of normal distributions with equal covariance matrices, was proposed by Anderson (1958), studied by John (1960, 1963), and has become known as John's Z statistic. Krzanowski (1982) extended this approach to mixed discrete and continuous variables. For more discriminant procedures in the mixture case, see Knoke (1982), Krzanowski (1975, 1979, 1980), and Tu and Han (1982).

Most of these classical classification rules allocate the individual to be classified to one of the populations if the ratio is less than a cut-off point c, and to the other otherwise. The cut-off point c is usually based on the probabilities of drawing an observation from the individual populations and the costs of misclassification. Associated with these procedures are the resulting misclassification probabilities. When, as in the problem of interest here, it is important to fix one of these probabilities of misclassification, the statistician will need to determine the cut-off point to allow this probability of misclassification to be prespecified.

When this probability is prespecified the problem then becomes one of testing a hypothesis. However, because of the setting of this problem we shall continue to refer to it as a classification problem. When the p-dimensional characteristic variable $\mathbf{V} \sim N_p(\boldsymbol{\mu}^{(0)}, \Sigma)$ for a population π_0, $\mathbf{V} \sim N_p(\boldsymbol{\mu}^{(1)}, \Sigma)$ for another population π_1, and $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\mu}^{(1)}$, Σ are unknown, Anderson (1973) and Kanazawa (1979) obtained the asymptotically normal expansion of the distribution of statistics W and Z respectively, which are used to find the cut-off point for a fixed value of the particular misclassification probability. In other cases (for example $\Sigma^{(0)} \neq \Sigma^{(1)}$ or V not normal) the asymptotic distribution of the classification statistics is, in general, unknown so that no hypothesis test is available.

In this report we determine a test of the classification hypothesis that satisfies the following requirements:

i) $\Sigma^{(0)}$ is not necessarily equal to $\Sigma^{(1)}$

ii) The p-dimensional discriminant variable may be a mixture of continuous and discrete variables

iii) The continuous variables need not be normally distributed.

Examples of continuous discriminants that are commonly used in the nuclear monitoring setting are ratios of amplitudes or spectra for different time windows and frequency bands of the observed seismogram. Earthquakes typically generate more shear energy than compressional energy, while explosions usually have much more compressional energy than shear. Since compressional waves propagate faster than shear elastic waves, this leads to larger relative amplitudes in different time windows for the two source types. Although explosive devices are expected to have more intrinsic high frequency content than earthquakes, explosions are usually shallower, in more anelastic materials than the deeper earthquakes, which tends to attenuate the high frequency content. As a result, spectral ratios of particular portions of the seismograms are useful discriminants in some regions of the world.

Some examples of categorical variables that are commonly used are presence of cepstral peaks, regional seismicity (high/low), location (off-shore/on-shore), depth (deep/shallow), and, in the context of associating mine blasts with a particular mine, day of the week.

The inability to treat a mixture of discrete and continuous variables rigorously in this setting has limited the application of many statistical classification methods in the past. This has led to rule-based approaches (Sereno and Wahl, 1993) which are somewhat ad hoc, artificial intelligence approaches (Baumgardt, et al, 1992), or inappropriate applications of linear discriminant functions or chi-squared tests. It is vital, however, for monitoring applications that these issues are all addressed with statistical rigor so that the error rates involved have meaning. The classification method proposed here satisfactorily addresses this problem by applying the bootstrap to the generalized likelihood ratio. Although this method is actually a test of hypothesis, it could just as well be used as a method for classification in the classical sense with the bootstrap being used to determine the probabilities of misclassification. For additional discussion of procedures for classifying seismic events see Shumway(1988).

In Section 2, we discuss the motivation for the proposed bootstrap likelihood ratio classification procedure, show how to construct the bootstrap likelihood ratio statistic, and explain how to determine the cut-off point for a desired misclassification probability. Section 3 is devoted to the application of the procedure to three cases. In *Example 1*, the bootstrap likelihood ratio statistic is shown to perform almost as well as the statistics W and Z which are specifically designed for *Example 1*, i.e. the case where two normal distributions with the same covariance matrix are considered. The bootstrap also performs quite well for both the normal case with different covariance matrices (*Example 2*) and the case of a mixture of continuous and discrete variates (*Example 3*), where, in either case, classical classification rules cannot control the probability of misclassification since their limiting distributions are unknown. In *Example 4* we apply the results developed here to some real seismic discriminant data and in Section 4 we present some concluding remarks.

2. Bootstrap Generalized Likelihood Ratio Test for Classification

2.1. Motivation

Let $\mathbf{V}' = (V_1, \ldots, V_p)$ be a p-dimensional random vector which is used to classify an individual into either population π_0 or population π_1. For $i = 0, 1$, let $f_i(\mathbf{v} \mid \theta^{(i)})$ be the probability or probability density function of $\mathbf{V}$ evaluated at $\mathbf{v}$, if $\mathbf{v}$ comes from population π_i, where $\theta^{(i)}$ is the set of unknown parameters. The components of V may be all discrete, all continuous, or mixture of discrete and continuous variables. In the mixed variables case, for example, let $\mathrm{V}' = (\mathrm{Y}, \mathrm{X})$ with $\mathrm{Y} = (\mathrm{Y}_1, \ldots, \mathrm{Y}_k)$ and $\mathrm{X} = \mathrm{X}_1, \ldots, \mathrm{X}_{p-k}$, where $\mathrm{Y}_1, \ldots, \mathrm{Y}_k$ are discrete and $\mathrm{X}_1, \ldots, \mathrm{X}_{p-k}$ are continuous. Suppose Y has the probability $f_{i,\mathbf{Y}}(\mathrm{Y}|\theta^{(i)}_{\mathbf{Y}})$ and the conditional probability density function of X given Y is $f_{i,\mathbf{X}|\mathbf{Y}}(\mathrm{X}|\theta^{(i)}_{\mathbf{X}|\mathbf{Y}}, \mathrm{Y})$. Then the joint probability density function of V in π_i is given by

$$f_i(\mathbf{v}|\theta^{(i)}) = f_{i,\mathbf{Y}}(\mathbf{y}|\theta^{(i)}_{\mathbf{Y}}) f_{i,\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\theta^{(i)}_{\mathbf{X}|\mathbf{Y}}, \mathbf{y}), \tag{1}$$

where $\boldsymbol{\theta}^{(i)} = \{\boldsymbol{\theta}^{(i)}_{\mathbf{Y}}, \boldsymbol{\theta}^{(i)}_{\mathbf{X}|\mathbf{Y}}\}$, $i = 0, 1$. See Olkin and Tate (1961) for the mixture of the multinomial and the multivariate normal distributions.

For any given classification rule, suppose that the region R_i is such that $\mathbf{v} \in R_i$ implies that $\mathbf{v}$ is classified as belonging to π_i. Further assume that $R_0 \cap R_1 = \emptyset$. The respective probabilities of misclassification are

$$P(1|0) = \int_{R_1} f_0(\mathbf{v} \mid \boldsymbol{\theta}^{(0)})\, d\mathbf{v}$$

$$P(0|1) = \int_{R_0} f_1(\mathbf{v} \mid \boldsymbol{\theta}^{(1)})\, d\mathbf{v},$$

where $d\mathbf{v} = dv_1 . . . dv_p$. The classical classification rules obtain the optimal regions R_0 and R_1 based on $f_0(\mathbf{v} \mid \boldsymbol{\theta}^{(0)}) \big/ f_1(\mathbf{v} \mid \boldsymbol{\theta}^{(1)})$ according to their classification principles (such as minimization of the total probability of misclassification, minimization of the total cost of misclassification, maximization of the posterior probability, minimax classification, etc.). However under any one of these classification principles, neither $P(1|0)$ nor $P(0|1)$ is fixed in advance at a certain value, which here we desire.

2.2. Bootstrapping the Log Likelihood Ratio Test Statistic

Suppose we have the training samples $\{\mathbf{v}^{(0)}_1, \mathbf{v}^{(0)}_2, . . . , \mathbf{v}^{(0)}_{N_0}\}$ of size N_0, and $\{\mathbf{v}^{(1)}_1, \mathbf{v}^{(1)}_2, . . . , \mathbf{v}^{(1)}_{N_1}\}$ of size N_1 from π_0 and π_1, respectively. A new observation whose value is $\mathbf{v}$ must be classified as from either π_0 or π_1. Now we employ a hypothesis-testing approach to classify $\mathbf{v}$. That is, the classification of $\mathbf{v}$ is accomplished by testing the hypothesis

$$\mathrm{H}_0\text{: } \mathbf{v}, \mathbf{v}^{(0)}_1, \mathbf{v}^{(0)}_2, \ldots, \mathbf{v}^{(0)}_{N_0} \in \pi_0\ ;\ \ \mathbf{v}^{(1)}_1, \mathbf{v}^{(1)}_2, \ldots, \mathbf{v}^{(1)}_{N_1} \in \pi_1$$

$$\mathrm{H}_1\colon\ \mathbf{v}_1^{(0)}, \mathbf{v}_2^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)} \in \pi_0\ ;\ \mathbf{v}, \mathbf{v}_1^{(1)}, \mathbf{v}_2^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)} \in \pi_1.$$

We use the generalized likelihood ratio method to construct a test. The likelihood of the two training samples is given by

$$L(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)} \mid \mathbf{v}_1^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)}, \mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)}) = \prod_{j=1}^{N_0} f_0(\mathbf{v}_j^{(0)} \mid \boldsymbol{\theta}^{(0)}) \prod_{j=1}^{N_1} f_1(\mathbf{v}_j^{(1)} \mid \boldsymbol{\theta}^{(1)}). \quad (2)$$

Consider now the new individual $\mathbf{v}$ to be classified. If this individual is included with the training sample from π_i, then an extra multiplying factor

$$\mathrm{L}_i(\boldsymbol{\theta}^{(i)} \mid \mathbf{v}) = f_i(\mathbf{v} \mid \boldsymbol{\theta}^{(i)})$$

must be incorporated in (2). The generalized likelihood ratio is therefore either unity or given by

$$LR = \frac{\sup_{\{\boldsymbol{\theta}^{(0)},\boldsymbol{\theta}^{(1)} \mid \mathrm{H}_0\}} \{L_0(\boldsymbol{\theta}^{(0)} \mid \mathbf{v})\ L(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)} \mid \mathbf{v}_1^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)}, \mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)})\ \}}{\sup_{\{\boldsymbol{\theta}^{(0)},\boldsymbol{\theta}^{(1)} \mid \mathrm{H}_1\}} \{L_1(\boldsymbol{\theta}^{(1)} \mid \mathbf{v})\ L(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)} \mid \mathbf{v}_1^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)}, \mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)})\}}$$

$$= \frac{L_0(\hat{\boldsymbol{\theta}}_0^{(0)} \mid \mathbf{v})\ L(\hat{\boldsymbol{\theta}}_0^{(0)}, \hat{\boldsymbol{\theta}}_0^{(1)} \mid \mathbf{v}_1^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)}, \mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)})}{L_1(\hat{\boldsymbol{\theta}}_1^{(1)} \mid \mathbf{v})\ L(\hat{\boldsymbol{\theta}}_1^{(0)}, \hat{\boldsymbol{\theta}}_1^{(1)} \mid \mathbf{v}_1^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)}, \mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)})}\ , \quad (3)$$

where $\hat{\boldsymbol{\theta}}_0^{(i)}$ is the Maximum Likelihood Estimator (MLE) of $\boldsymbol{\theta}^{(i)}$ under H_0 and $\hat{\boldsymbol{\theta}}_1^{(i)}$ is the MLE of $\boldsymbol{\theta}^{(i)}$ under H_1, $i = 0, 1$. Now let $\lambda = \log(LR)$. It intuitively follows that small values of λ provide evidence against H_0 and thus the generalized likelihood ratio test is to reject H_0 if $\lambda \le \lambda_\alpha$, where λ_α is chosen to provide a size α test.

Let $P(\lambda \le \lambda_\alpha \mid \mathrm{H}_0)$ denote the size of the Type I error and $P(\lambda > \lambda_\alpha \mid \mathrm{H}_1)$ denote the size of the Type II error for a constant λ_α. Then $P(\lambda \le \lambda_\alpha \mid \mathrm{H}_0)$ is the probability of misclassification $P(1|0)$, and $P(\lambda > \lambda_\alpha \mid \mathrm{H}_1)$ is the probability of misclassification $P(0|1)$ when R_0 and R_1 are defined in terms of λ_α. Therefore we can construct a classification rule which can control one of the probabilities of misclassification by fixing the size of the test if we know the distribution of $\lambda(\mathbf{V}, \mathbf{V}_1^{(0)}, \ldots, \mathbf{V}_{N_0}^{(0)}, \mathbf{V}_1^{(1)}, \ldots, \mathbf{V}_{N_1}^{(1)})$. In most cases it is difficult to obtain the exact distribution of the test statistic λ. The distribution, however, can be approximated by employing the bootstrap method (Efron 1979, 1982).

Since the form of the probability density function is assumed known, the bootstrap samples can be obtained from the estimated density function. This is called the parametric bootstrap (Efron 1979), and we employ it in this study. We have examined the use of the nonparametric approach of resampling with replacement from the training samples, and for the training samples of size 25 or larger, this nonparametric bootstrapping yielded similar results to those reported here.

The likelihood ratio statistic for the test of the null hypothesis H_0 versus the alternative H_1 can be parametrically bootstrapped as follows. Given the training samples $\{\mathbf{v}_j^{(0)}\}_{j=1}^{N_0}$, $\{\mathbf{v}_j^{(1)}\}_{j=1}^{N_1}$, bootstrap samples $\{\mathbf{v}_j^{*(0)}\}_{j=1}^{N_0+1}$, $\{\mathbf{v}_j^{*(1)}\}_{j=1}^{N_1}$ are generated randomly from $f_0(\mathbf{v} \mid \hat{\boldsymbol{\theta}}_1^{(0)})$ and $f_1(\mathbf{v} \mid \hat{\boldsymbol{\theta}}_0^{(1)})$, respectively, where $\hat{\boldsymbol{\theta}}_1^{(0)}$ and $\hat{\boldsymbol{\theta}}_0^{(1)}$ are obtained from the original samples $\{\mathbf{v}_j^{(0)}\}_{j=1}^{N_0}$ and $\{\mathbf{v}_j^{(1)}\}_{j=1}^{N_1}$, respectively. The value of λ, to be denoted λ^*, is computed for the bootstrap samples by substituting $\mathbf{v}_{N_0+1}^{*(0)}$, $\{\mathbf{v}_1^{*(0)}, \ldots, \mathbf{v}_{N_0}^{*(0)}\}$, $\mathbf{v}_1^{*(1)}, \ldots, \mathbf{v}_{N_1}^{*(1)}\}$ for $\mathbf{v}$, $\{\mathbf{v}_1^{(0)}, \ldots, \mathbf{v}_{N_0}^{(0)}\}$, $\{\mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{N_1}^{(1)}\}$ in (3), respectively. This process is repeated independently B times, and the replicated values of λ^*, $\{\lambda_j^*\}_{j=1}^{B}$, evaluated from the successive bootstrap samples, can be used to assess the true null distribution of λ. In particular, the αth empirical quantile of $\{\lambda_j^*\}_{j=1}^{B}$, denoted by λ_α^*, will essentially approach λ_α, the true critical value for the test of size α, for large N_0 and N_1 as B tends to infinity. (See Bickel and Freedman (1981) for some asymptotic theory on the quantile process for the bootstrap.). Thus we use λ_α^* as a critical value for the test of size α. Therefore, we allocate $\mathbf{v}$ to π_1 if $\lambda \leq \lambda_\alpha^*$, and allocate $\mathbf{v}$ to π_0, otherwise.

McLachlan (1987) showed the relationship between λ_α^* and the bootstrap replication size B for the specified test size α. In general, given a set of B order statistics from a population, the probability that a randomly selected member from the population is less than or equal to the jth order statistic is $j/(B+1)$. Thus, if $\alpha = j/(B+1)$, then λ^*_α is the jth smallest value of $\{\lambda^*_i\}_{i=1}^{B}$, i.e. if $\alpha = 0.05$ and $B = 299$ then λ^*_α is the 15th smallest value of $\{\lambda^*_i\}_{i=1}^{299}$.

3. Applications

The bootstrap generalized likelihood ratio test proposed here allows the p-dimensional characteristic variable $\mathbf{V}$ to be discrete, continuous, or a combination of discrete and continuous variables, and its probability or probability density function $f_i(\mathbf{V} \mid \boldsymbol{\theta}^{(i)})$ for π_i is assumed to be known except for the value of the parameter $\boldsymbol{\theta}^{(i)}$, $i =$ 0, 1. It can therefore be applied to the classification problem in each of these cases when one needs to control one of the probabilities of misclassification. As we will see, the bootstrap generalized likelihood ratio test essentially achieves the required probability of misclassification for even a moderate size sample. Throughout, we assume that we have random samples $\{\mathbf{v}_j^{(0)}\}_{j=1}^{N_0}$ from π_0, and $\{\mathbf{v}_j^{(1)}\}_{j=1}^{N_1}$ from π_1.

In the following four examples we consider four distinct scenarios. In the first example we consider the simple case where the observations are all normal with equal covariances. Of course this case is well established, but we consider it to demonstrate that very little is lost by using the bootstrap rather than the exact distribution. In *Example 2*, we continue to assume normality but drop the assumption of equal covariances. In this case the bootstrap is necessary in order to determine the proper critical point. However, it is not necessary to bootstrap the likelihood ratio, but instead one could bootstrap the quadratic discriminant function, Q. This example demonstrates that these two bootstrap approaches yield essentially the same result. In *Example 3* we consider a mixture of normal and binomial variates where, to our knowledge, no alternative to the method introduced here is available. Finally, in *Example 4* we consider a set of real data which is treated as a mixture of normal and multinomial data.

Example 1: Normal Distributions with Equal Covariance Matrix

Suppose that $f_i(\mathbf{v} \mid \boldsymbol{\theta}^{(i)})$ is the density function for $\mathrm{N}_p(\boldsymbol{\mu}^{(i)}, \Sigma^{(i)})$ with $\Sigma^{(0)} = \Sigma^{(1)}$, $(= \Sigma)$, where $\boldsymbol{\theta}^{(i)} = (\boldsymbol{\mu}^{(i)},\Sigma)$. Replacing the unknown parameters in $f_0(\mathbf{v} \mid (\boldsymbol{\mu}^{(0)}, \Sigma))/f_2(\mathbf{v} \mid (\boldsymbol{\mu}^{(1)}, \Sigma))$ by their estimates leads to the well-known Anderson's W statistic (A2) given in the appendix. The likelihood ratio (A3) is characterized by John's Z statistic (A4). On the other hand, the log likelihood ratio statistic, λ, is given in (A4) and is obtained directly by taking the log of the expression (A3) and dividing it by a constant. The monotonic relationship between Z and λ is obvious. If the values of W, Z, and λ are greater than their cut-off points, then π_0 is favored for $\mathbf{v}$, and π_1 is preferred otherwise.

Now we want to choose the cut-off point so that one probability of misclassification is controlled. Let α be the desired $P(1|0)$. Anderson (1973) has obtained from the asymptotic normal distribution of W, the following approximate cut-off point W_α, which attains the desired probability α to within $\mathrm{O}(N^{-2})$. For large N_0 and N_1,

$$W_\alpha = \frac{1}{2} D^2 + D\left[u_0 - \frac{1}{N_0}\left(\frac{p-1}{D} - \frac{1}{2} u_0\right) + \frac{1}{N}\left(\left(p-\frac{3}{4}\right) u_0 + \frac{1}{4} u_0^3\right)\right],$$

where $N = N_0 + N_1 - 2$, $D = \sqrt{(\bar{\mathbf{v}}^{(0)} - \bar{\mathbf{v}}^{(1)})'\mathrm{S}^{-1}(\bar{\mathbf{v}}^{(0)} - \bar{\mathbf{v}}^{(1)})}$, u_0 is such that $\Phi(u_0) = \alpha$, and $\Phi\,(\cdot)$ is the cumulative $N(0,\,1)$ distribution function. Kanazawa (1979) has obtained the asymptotic cut-off point Z_α for the Z statistic. For large N_0 and N_1,

$$Z_\alpha = \frac{1}{2} D^2 + D\,\Big[u_0 + \frac{1}{2N_0 D}\left(u_0^2 + Du_0 - (p-1)\right)$$

$$-\frac{1}{2\,N_1\,D}\left(u_0^2 + 2Du_0 + (p-1) + D^2\right) + \frac{1}{4N}\left(u_0^3 + (4p-3)u_0\right)\Big],$$

where D and u_0 are the same as above.

Instead of deriving the limiting distribution, the cut-off point λ_α^* of the bootstrap log likelihood ratio statistic λ is obtained by the parametric bootstrap procedure described in Section 2.2. Using the MLEs of $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\mu}^{(1)}$, and Σ from the training samples $\{\mathbf{v}_j^{(\mathrm{i})}\}_{j=1}^{N_i}$, $i = 0,\ 1$, bootstrap samples $\{\mathbf{v}_j^{*(0)}\}_{j=1}^{N_0+1}$, $\{\mathbf{v}_j^{*(1)}\}_{j=1}^{N_1}$ are generated from a $\mathbf{N}_{\boldsymbol{p}}(\overline{\mathbf{v}}^{(0)},\ \mathbf{A}/(N_0 + N_1))$ and a $\mathbf{N}_{\boldsymbol{p}}(\overline{\mathbf{v}}^{(1)},\ \mathbf{A}/(N_0 + N_1))$, respectively where $\mathbf{A}$ is defined in the appendix. We compute the value of the log likelihood ratio statistic, λ^* corresponding to (A4), for the bootstrap samples by replacing $\mathbf{v}$, $\overline{\mathbf{v}}^{(0)}$, $\overline{\mathbf{v}}^{(1)}$, S by $\mathbf{v}_{N_0+1}^{*(0)}$, $\overline{\mathbf{v}}^{*(0)}$, $\overline{\mathbf{v}}^{*(1)}$, S^*, respectively, where $\overline{\mathbf{v}}^{*(\mathrm{i})} = \sum_{j=1}^{N_i} \mathbf{v}_j^{*(i)}/N_i$, i=0, 1, and S^* is calculated according to (A1) for the bootstrap samples. This process is repeated independently B times. Then λ_α^* is the αth empirical quantile of $\{\lambda_j^*\}_{j=1}^B$, where $\{\lambda_j^*\}_{j=1}^B$ are the values of λ^* evaluated from the successive bootstrap samples.

For given α, let $P_W(1|0)$, $P_Z(1|0)$ and $P_\lambda(1|0)$ be the probabilities that the new individual is misclassified into π_1 by the statistics W, Z and λ using the cut-off points W_α, Z_α, λ_α^*, respectively. Then $P_W(1|0) = P(W \leq W_\alpha|\pi_0)$, $P_Z(1|0) = P(Z \leq Z_\alpha|\pi_0)$, and $P_\lambda(1|0) = P(\lambda \leq \lambda_\alpha^* \mid \pi_0)$. We will examine how close $P_W(1|0)$, $P_Z(1|0)$ and $P_\lambda(1|0)$ are to the desired misclassification probability, $\alpha = P(1|0)$, for the normal distributions with equal covariance matrix by Monte Carlo method. We generate two sets of random samples $\left\{\mathbf{v}_i, \{\mathbf{v}_{ij}^{(0)}\}_{j=1}^{N_0}\right\}_{i=1}^M$, $\left\{\{\mathbf{v}_{ij}^{(1)}\}_{j=1}^{N_1}\right\}_{i=1}^M$ from $\mathbf{N}_2(\boldsymbol{\mu}^{(0)},\ \Sigma)$ and $\mathbf{N}_2(\boldsymbol{\mu}^{(1)},\ \Sigma)$, respectively, where

$$\boldsymbol{\mu}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \boldsymbol{\mu}^{(1)} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \text{and} \sum = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

For each $i = 1, 2, \ldots, M$, we obtain the values of the statistics W, Z, λ, say W_i, Z_i, λ_i, using $\{\mathbf{v}_i,\ \{\mathbf{v}_{ij}^{(0)}\}_{j=1}^{N_0},\ \{\mathbf{v}_{ij}^{(1)}\}_{j=1}^{N_1}\}$, and compare them to their corresponding critical values $W_{i\alpha}$, $Z_{i\alpha}$, $\lambda^*_{i\alpha}$ for a fixed α. $B = 499$ bootstrap samples are used for $\lambda^*_{i\alpha}$. Then $P_W(1|0)$, $P_Z(1|0)$ and $P_\lambda(1|0)$ are estimated by the proportion of times that the value of the statistic is less than or equal to its critical value among M trials. Since $\hat{P}_W(1|0)$ is the usual estimate of a proportion, its standard deviation (s.d.) is estimated by $\sqrt{\hat{P}_W(1|0)(1- \hat{P}_W(1|0))/M}$. The standard deviation estimates of $\hat{P}_Z(1|0)$ and $\hat{P}_\lambda(1|0)$ are obtained similarly. The first portion of Table 1 shows the estimates of the probability of misclassification with their standard deviations (s.d.) for the different sample sizes with $\alpha = 0.05$, $M = 10{,}000$. The results for $\hat{P}_W(1|0)$ and $\hat{P}_Z(1|0)$ are identical when $N_0 = N_1 = 25$ since $Z = (N_0/(N_0 + 1))\,W$ for $N_0 = N_1$. Although for the sample sizes considered, the bootstrap estimate does not attain the same precision as the W or Z statistic's estimate, it is clearly competitive.

Table 1. The estimates of the probability of misclassification, $P(1|0) = 0.05$, and the estimates of the power, $P(1|1)$

$\hat{P}_W(1\|0)$	$\hat{P}_Z(1\|0)$	$\hat{P}_\lambda(1\|0)$
	$N_0 = N_1 = 25$	
0.054 (0.002)	0.054 (0.002)	0.061 (0.002)
	$N_0 = 30,\ N_1 = 45$	
0.055 (0.002)	0.055 (0.002)	0.060 (0.002)
$\hat{P}_W(1\|1)$	$\hat{P}_Z(1\|1)$	$\hat{P}_\lambda(1\|1)$
	$N_0 = 30,\ N_1 = 45$	
0.726 (0.004)	0.725 (0.004)	0.736 (0.004)

Now we compare the powers, $P(1|1)$, for W, Z and λ. Random samples $\{\{\mathbf{v}_{ij}^{(0)}\}_{j=1}^{N_0}\}_{i=1}^{M}$, $\{\mathbf{v}_i, \{\mathbf{v}_{ij}^{(1)}\}_{j=1}^{N_1}\}_{i=1}^{M}$ are generated from $\mathrm{N}_2(\boldsymbol{\mu}^{(0)}, \Sigma)$ and $\mathrm{N}_2(\boldsymbol{\mu}^{(1)}, \Sigma)$, respectively with the same parameters as above. The power estimates for W, Z and λ, $\hat{P}_W(1|1)$, $\hat{P}_Z(1|1)$ and $\hat{P}_\lambda(1|1)$, are obtained in the same way as for $\hat{P}_W(1|0)$, $\hat{P}_Z(1|0)$ and $\hat{P}_\lambda(1|0)$, respectively. For $\alpha = 0.05$, $N_0 = 30$, $N_1 = 45$, $M = 10{,}000$ and $B = 499$, the power estimates are similar to each other with the bootstrap being slightly better (undoubtedly, due to the slightly larger critical region) as shown in the second portion of Table 1.

Example 2: Normal Distributions with Unequal Covariance Matrices

Let π_0 and π_1 be $\mathrm{N}_p(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})$ and $\mathrm{N}_p(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$ with $\boldsymbol{\mu}^{(0)} \neq \boldsymbol{\mu}^{(1)}$ and $\Sigma^{(0)} \neq \Sigma^{(1)}$. When the parameters are unknown, a classical classification rule known as the quadratic discriminant function is obtained by taking the log after substituting estimates, $\overline{\mathrm{v}}^{(0)}$, $\overline{\mathrm{v}}^{(1)}$, $\mathrm{S}^{(0)}$ and $\mathrm{S}^{(1)}$ of $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\mu}^{(1)}$, $\Sigma^{(0)}$, and $\Sigma^{(1)}$ into the ratio of the two multivariate normal probability density functions, $f_0(\mathrm{v}\ |\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})\ /\mathrm{f}_1(\mathrm{v} \mid \boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$. The quadratic discriminant function Q is given in (A5), and $\mathbf{v}$ is classified to π_0 if $Q > 0$ and to π_1 otherwise. The probabilities of misclassification of Q are difficult to control since even its limiting distribution is unknown.

Following the hypothesis-testing approach of (2), the MLEs of $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\mu}^{(1)}$, $\Sigma^{(0)}$, $\Sigma^{(1)}$ under H_0 and H_1 are given in the Appendix. The log likelihood ratio statistic, λ, is given in (A6), and to evaluate the cut-off point λ_α^*, for the desired probability of misclassification, $P(1|0) = \alpha$, we generate bootstrap samples $\{\mathbf{v}_j^{*(0)}\}_{j=1}^{N_0+1}$, $\{\mathbf{v}_j^{*(1)}\}_{j=1}^{N_1}$ from a $\mathbf{N}_{\boldsymbol{p}}(\overline{\mathbf{v}}^{(0)}, \mathbf{A}^{(0)}/N_1)$ and a $\mathbf{N}_{\boldsymbol{p}}(\overline{\mathbf{v}}^{(1)}, \mathbf{A}^{(1)}/\mathrm{N}_2)$, respectively. Following the same bootstrap procedure as in *Example 1,* the αth empirical quantile λ_α^* is obtained from the values of the log likelihood ratio statistic λ for the successive bootstrap samples. The bootstrap generalized likelihood ratio classification rule with misclassification probability $P(1|0) = \alpha$ is, therefore, to assign $\mathbf{v}$ to π_1 if $\lambda(\mathbf{v}) \leq \lambda^*_\alpha$, and to π_0, otherwise.

Consider two bivariate normal distributions $\mathrm{N}_2(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})$, $\mathrm{N}_2(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$, where

$$\boldsymbol{\mu}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}^{(1)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \Sigma^{(0)} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \text{ and } \Sigma^{(1)} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}.$$

Suppose we apply the Q statistic for classification using the usual classification rule, *i.e.* $\mathbf{v}$ is classified to π_0 if $Q > 0$ and to π_1 otherwise. The probability of misclassification of interest, i.e. $P_Q(1|0)$ is $P(Q \leq 0 \,|\pi_0)$. In order to determine the probability of this classification error we conduct a simulation. We generate $\{\mathbf{v}_i, \{\mathbf{v}_{ij}^{(0)}\}_{j=1}^{N_0}\}_{i=1}^{M}$, and $\{\mathbf{v}_{ij}^{(1)}\}_{j=1}^{N_1}\}_{i=1}^{M}$ from $\mathrm{N}_2(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})$ and $\mathrm{N}_2(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$, respectively. We obtain the Q statistics for $\{\mathbf{v}_i, \{\mathbf{v}_{ij}^{(0)}\}_{j=1}^{\mathrm{N}_0}\}$, $\{\mathbf{v}_{ij}^{(1)}\}_{j=1}^{N_1}$, $i = 1, 2, \ldots, M$, and denote these Q_1, $Q_2, \ldots, Q_M$. Then $P_Q(1|0)$ is estimated by $\hat{P}_Q(1|0)$ which is the proportion of Q_i values that are less than or equal to zero. $\hat{P}_Q(1|0)$ (with its standard deviation) is 0.274 (0.004) for $N_0 = 100$, $N_1 = 150$ and $M = 10{,}000$. When it is important to keep the probability of misclassification $P_Q(1|0)$ small, an error this large may be unacceptable, resulting in the need for the method we are describing.

Now we consider the log likelihood ratio statistic λ. First, we would like to know how well the parametric bootstrap procedure approximates the true null distribution of λ. Since the true null distribution of λ is not known, we generate samples $\{\mathbf{v}_i, \{\mathbf{v}_{ij}^{(0)}\}_{j=1}^{N_0}\}_{i=1}^{M}$ from $\mathrm{N}_2(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})$ and $\{\{\mathbf{v}_{ij}^{(1)}\}_{j=1}^{N_1}\}_{i=1}^{M}$ from $\mathrm{N}_2(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$ with $M = 100{,}000$. Applying $\{\mathbf{v}_i, \{\mathbf{v}_{ij}^{(0)}\}_{j=1}^{N_0}, \{\mathbf{v}_{ij}^{(1)}\}_{j=1}^{N_1}\}_{i=1}^{M}$ to (A6), we can obtain $\{\lambda_i\}_{i=1}^{M}$. The true null cumulative distribution function (cdf) of λ is approximated by the empirical cdf using $\{\lambda_i\}_{i=1}^{M}$ for $(N_0, N_1) = (10, 15)$, $(N_0, N_1) = (30{,}45)$, and $(N_0, N_1) = (100, 150)$. The true critical value λ_α is approximated by -1.900, -1.504, -1.353 respectively. These are the αth quantiles of $\{\lambda_i\}_{i=1}^{M}$ where $\alpha = 0.05$ for $(N_0, \mathrm{N}_1) = (10, 15)$, $(N_0, N_1) = (30,$

45), and $(N_0, N_1) = (100, 150)$, respectively. In this simulation, $B = 299$ is used for the bootstrap replication size because of computer-time constraints. Our investigation indicates that the results using $B = 299$ and $B = 499$ are similar.

For a set of random samples $\{\mathbf{v}, \{\mathbf{v}_j^{(0)}\}_{j=1}^{N_0}\}, \{\mathbf{v}_j^{(1)}\}_{j=1}^{N_1}\}$ under H_0 with $(N_0, N_1) = (10, 15)$, $(N_{0,} N_1) = (30, 45)$, and $(N_0, N_1) = (100, 150)$, the empirical null distribution of the bootstrap log likelihood statistic using $\{\lambda_j^*\}_{j=1}^{B}$ with $B = 299$ is also plotted around the true null cdf in Figure 1. Inspection of this figure shows that the bootstrap null distribution approximates the true null distribution of the log likelihood ratio statistic quite well as the sample sizes increase and does surprisingly well for small samples.

Even though the null distribution of the Q statistic is unknown, the cut-off point, Q_{α}, for misclassification probability, $P(1|0) = \alpha$, can be approximated by the same parametric bootstrap procedure as for λ. That is, we evaluate the Q statistic for B successive bootstrap samples and call them $Q_1^*, Q_2^*, \ldots, Q_B^*$. Then Q_α is approximated by Q_α^*, the αth empirical quantile of $\{Q_j^*\}_{j=1}^{B}$. Therefore, one can allocate $\mathbf{v}$ to π_1 if $Q \leq Q_\alpha^*$, and allocate $\mathbf{v}$ to π_0, otherwise.

With the same simulation data used to get $\hat{P}_Q(1|0) = 0.274$ above, $\hat{P}_{QB}(1|0)$ (s.d.), the estimate of a fixed $P(1|0) = 0.05$ by the parametrically bootstrapped Q statistic QB, is 0.050 (0.002) for $B = 499$. $\hat{P}_{\lambda}(1|0)$ (s.d.) of the bootstrapped λ, i.e. λ^*, is 0.049 (0.002) for the same bootstrap samples as for QB. Both bootstrap estimates are close to the true fixed misclassification probability $P(1|0) = 0.05$.

To further compare the two tests we now investigate their respective powers, $P(1|1)$, for different parameter values. Consider bivariate normal distributions $\mathbf{N}_2(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})$ and $\mathbf{N}_2(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$. Let ρ_0 and ρ_1 be the correlation coefficient for $\mathbf{N}_2(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})$ and $\mathbf{N}_2(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$ respectively. We assume that $\rho_0 = 0.5$, $\rho_1 = -0.5$ and that both distributions have the same marginal variances, $\sigma_1^2 = 1$ and $\sigma_2^2 = 1$. That is,

$$\sum{}^{(0)} = \begin{pmatrix} \sigma_1^2 & \rho_0\sigma_1\sigma_2 \\ \rho_0\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix},$$

$$\sum{}^{(1)} = \begin{pmatrix} \sigma_1^2 & \rho_1\sigma_1\sigma_2 \\ \rho_1\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix},$$

For $\boldsymbol{\mu}^{(0)} = (0,\ 0)'$, we examine the power, $P(1|1)$ of the bootstrap Q statistic and the bootstrap λ statistic at $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(0)} + \Delta(\sigma_1,\ \sigma_2)'$, $\Delta = 1,\ 2,\ 3$, for small samples ($N_0 = 10$, $N_1 = 15$) and for large samples ($N_0 = 100$, $N_1 = 150$). For each $\Delta = 1,\ 2,\ 3$ under H_1, we randomly generate $\{\{\mathbf{v}_{ij}^{(0)}\}_{j=1}^{N_0}\}_{i=1}^{M}$ from $\mathrm{N}_2(\boldsymbol{\mu}^{(0)},\ \Sigma^{(0)})$ and $\{\mathbf{v}_i,\ \{\mathbf{v}_{ij}^{(1)}\}_{j=1}^{N_1}\}_{i=1}^{M}$ from $N_2(\boldsymbol{\mu}^{(1)},\ \Sigma^{(1)})$ with $N_0 = 10$, $N_1 = 15$ and $M = 10{,}000$. For each $i = 1, \ldots, M$ and for $\alpha = 0.05$, $\{\mathbf{v}_i,\ \{\mathbf{v}_{ij}^{(0)}\}_{j=1}^{N_0},\ \{\mathbf{v}_{ij}^{(1)}\}_{j=1}^{N_1}\}$ is used for the parametric bootstrap to obtain the cut-off points Q_α^* and λ_α^* for QB and λ, respectively. The bootstrap replication size B used here is 499. Then the power estimate $\hat{P}_{QB}(1|1)$ for QB is the proportion of times that the Q statistic value is less than or equal to Q_α^* out of M trials. The power estimate $\hat{P}_\lambda(1|1)$ for the bootstrap λ is obtained similarly. $\hat{P}_{QB}(1|1)$ and $\hat{P}_\lambda(1|1)$ are listed along with those for large samples ($N_0 = 100$, $N_1 = 150$) in Table 2.

Table 2. Power comparison between the bootstrap λ and the bootstrap Q (QB) with $B = 499$. Entry is power estimate with its standard deviation.

	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$
		$N_0 = 10,\ N_1 = 15$	
$\hat{P}_\lambda(1\|1)$	0.310 (0.0046)	0.815 (0.0039)	0.992 (0.0009)
$\hat{P}_{QB}(1\|1)$	0.302 (0.0046)	0.795 (0.0040)	0.990 (0.0010)
		$N_0 = 100,\ N_1 = 150$	
$\hat{P}_\lambda(1\|1)$	0.375 (0.0048)	0.884 (0.0032)	0.999 (0.0003)
$\hat{P}_{QB}(1\|1)$	0.376 (0.0048)	0.884 (0.0032)	0.999 (0.0003)

In this simulation, the bootstrap λ has slightly higher power than the bootstrap Q for small samples, but there is little difference for large samples.

Example 3: Mixture of Categorical and Continuous Variables

In this example we consider a mixture of continuous and discrete variates. Of the discriminant functions in the previous sections, only the λ statistic has been studied for this case. Suppose the variable **V** is a mixture of discrete and continuous variables. Let $\mathbf{V}' = (\boldsymbol{Z}, \mathrm{X})$ with $\boldsymbol{Z} = (Z_1, \ldots, Z_r)$ and $\mathrm{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_q)$ where $Z_1, \ldots, Z_r$ are discrete and $\mathrm{X}_1, \ldots, \mathrm{X}_q$ are continuous, r and q are positive integers. Suppose further the jth discrete variable Z_j has k_j categories, $j = 1, \ldots, r$. Then the vector of discrete variables $\boldsymbol{Z}$ may be expressed as a multinomial random variable $\boldsymbol{Y}' = (Y_1, \ldots, Y_k)$, where $Y_m = 0$ or 1, $m = 1, \ldots, k$, $\sum_{m=1}^{k} Y_m = 1$, and $k = \prod_{j=1}^{r} k_j$. Thus, each distinct pattern of $\boldsymbol{Z}$ defines a multinomial cell of $\boldsymbol{Y}$ uniquely. It is assumed that the probability of obtaining an observation in cell m for π_i is $p_m^{(\mathrm{i})}$, $(0 \leq p_m^{(\mathrm{i})} \leq 1,\ \sum_{m=1}^{k} p_m^{(i)} = 1)$, $i = 0, 1$. Then the joint probability density function of **V** in π_i is given by (1), where $\boldsymbol{\theta}_{\mathbf{Y}}^{(i)\prime} = (p_1^{(i)}, \ldots, p_{k\text{-}1}^{(i)})$ and $\boldsymbol{\theta}_{\mathbf{X}|\boldsymbol{Y}}^{(i)}$ is the set of parameters of **X** given **Y**.

For the population π_i, the conditional pdf of **X** given **Y**, $f_{i,\mathbf{X}|\mathbf{Y}}(\mathrm{X} \mid \mathrm{Y})$, may be of any proper type depending on the relationship between **X** and **Y**. Following Olkin and Tate (1961), for this example we assume that **X** has a conditional multivariate normal distribution with mean $\boldsymbol{\mu}_m^{(i)}$ given **Y** belonging to cell m and common covariance matrix $\Sigma^{(i)}$ in all cells. If **Y** belongs to cell m, i.e., if $\mathbf{Y} = (\mathrm{Y}_1, \cdots, \mathrm{Y}_{m-1}, \mathrm{Y}_m, \mathrm{Y}_{m+1}, \cdots, \mathrm{Y}_k) = (0, \cdots, 0, 1, 0, \cdots, 0)$, then $f_{i,\mathbf{Y}}(\mathrm{Y} \mid \boldsymbol{\theta}_{\mathbf{Y}}^{(i)})$ and $f_{i,\mathbf{X}|\mathbf{Y}}(\mathrm{X} \mid \boldsymbol{\theta}_{\mathbf{X}|\mathbf{Y}}^{(i)}, \mathrm{Y})$ of (1) are given as follows:

$$f_{i,\mathbf{Y}}(\mathrm{Y} | \boldsymbol{\theta}_{\mathbf{Y}}^{(i)}) = p_m^{(i)}$$

$$f_{i,\mathbf{X}|\mathbf{Y}}(\mathbf{X}|\theta^{(i)}_{\mathbf{X}|\mathbf{Y}}, \mathbf{Y}) = (2\pi)^{-q/2} \,|\, \Sigma^{(i)}|^{-1/2} exp\{-(1/2)(\mathbf{x}-\mu^{(i)}_m)'(\Sigma^{(i)})^{-1}(\mathbf{x}-\mu^{(i)}_m)\}.$$

Let the jth member of the training sample, $\{\mathbf{v}^{(i)}_1, \mathbf{v}^{(i)}_2, \ldots, \mathbf{v}^{(i)}_{N_i}\}$ from π_i be denoted by $\{\mathbf{v}^{(i)\prime}_j = (\mathbf{y}^{(i)}_j, \mathbf{x}^{(i)}_j)\}$, where $\mathbf{y}^{(i)}_j$ is the vector of binary variables obtained from the discrete components $\mathbf{z}$ of $\mathbf{v}^{(i)}_j$, and $\mathbf{x}^{(i)}_j$ is the vector of continuous variables. Let $n^{(i)}_m$ denote the number of individuals of the training sample from π_i that fall in cell m defined by $\mathbf{Y}$. Then $N_i = \sum^k_{m=1} n^{(i)}_m$, $i = 0, 1$. The likelihood of the two training samples is given by

$$L = \prod_{i=0}^{1} [\{\prod_{m=1}^{k} (p^{(i)}_m)^{n^{(i)}_m}\}\{(2\pi)^q \,|\Sigma^{(i)}|\}^{\frac{-N_i}{2}}$$

$$\cdot\ exp\{-\frac{1}{2} \sum_{j=1}^{N_i}(\mathbf{x}^{(i)}_j - \mu_{ij})' \,(\Sigma^{(i)})^{-1}\, (\mathbf{x}^{(i)}_j - \mu_{ij})\}] \,, \qquad (4)$$

where μ_{ij} takes the value $\mu^{(i)}_m$ if $\mathbf{y}^{(i)}_j$ falls in the mth cell, $m = 1, \ldots, k$.

Consider now the new individual $\mathbf{v}$ to be classified, and suppose that the discrete components place it into cell l. If this individual is included with the training sample from π_i, then an extra multiplying factor

$$L^{(i)}_l = (2\pi)^{-q/2} \,|\Sigma^{(i)}|^{-1/2} p^{(i)}_l \; exp\{-\tfrac{1}{2}(\mathbf{x}-\mu^{(i)}_l)' \,(\Sigma^{(i)})^{-1}(\mathbf{x}-\mu^{(i)}_l)\}$$

must be incorporated in (4) to construct the generalized likelihood ratio test statistic of (3). $\mathbf{x}^{(i)}_j$ must belong to one of k subgroups corresponding to the conditional distributions depending on the value of $\mathbf{y}^{(i)}_j$ for $j = 1, \ldots, N_i$, $i = 0, 1$. Let $\mathrm{x}^{(i)}_{sm}$ be the sth member of mth subgroup of the continuous variable measurements whose discrete covariates fall in the mth cell. Then any element of $\{\mathbf{x}^{(i)}_j\}^{N_i}_{j=1}$ belongs to one of k subgroups $\{\{\mathrm{x}^{(i)}_{sm}\}^{n^{(i)}_m}_{s=1}\}^k_{m=1}$ where, of course, some of the $n^{(i)}_m$ could be zero. Hence we can rewrite the exponent of (4) as

$$-\frac{1}{2}\sum_{m=1}^{k} \left(\sum_{s=1}^{n_m^{(i)}}(\mathbf{x}_{sm}^{(i)} - \boldsymbol{\mu}_m^{(i)})' \, (\Sigma^{(i)})^{-1}(\mathbf{x}_{sm}^{(i)} - \boldsymbol{\mu}_m^{(i)})\right).$$

The MLEs under H_0 and H_1 are given in the appendix, and the log likelihood ratio statistic is given in (A7).

Krzanowski (1982) considered a similar likelihood ratio statistic when $\Sigma^{(0)} = \Sigma^{(1)}$, $(= \Sigma)$, and $\boldsymbol{\mu}_m^{(i)}$ and Σ are estimated by a second-order regression model of $\mathbf{X}$ on $\mathbf{Y}$. Then he allocated a new individual to π_0 if his likelihood ratio statistic is greater than or equal to 1 and to π_1 otherwise. He did not consider the problem when it is desired to control one of the misclassification errors.

We investigate the performance of the bootstrap log likelihood ratio test by examining the power with a simulation. We consider a simple situation in which we have a discrete variable from a Bernoulli(p) distribution and an independent continuous variable distributed $N(\mu, \sigma^2)$. For $i = 0, 1$, let $\{\mathbf{v}_j^{(i)} = (z_j^{(i)}, x_j^{(i)})'\}_{j=1}^{N_i}$ be a random sample from π_i, where $z_j^{(i)} \sim$ Bernoulli(p_i) and $x_j^{(i)} \sim N(\mu_i, \sigma_i^2)$. Let $\mathbf{v} = (z, x)'$ be a new observation to be classified where $z \sim$ Bernoulli(p_1) and $x \sim N(\mu_1, \sigma_1^2)$.

We examine the power of the bootstrap λ, $\mathrm{P}_\lambda(1|1)$, for different parameter values. We set $p_0 = 0.1$, $\mu_0 = 0$, $\sigma_0 = 0.5$, and $\sigma_1 = 1$. For $p_1 = 0.9$, 0.7, and 0.5, the estimate of $P_\lambda(1|1)$ is obtained for $\mu_1 = 0.5 + \Delta\sigma_1$ where $\Delta = \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$. The power estimate, $\hat{P}_\lambda(1|1)$, is the proportion of times that the λ statistic value is less than or equal to λ_α^* out of 2000 trials, where λ_α^* is the bootstrap cut-off point at α significance level. With $N_0 = N_1 = 50$, $B = 299$, and $\alpha = 0.05$, these power estimates are plotted in Figure 2. As the separation between μ_0 and μ_1 increases, the power of the bootstrap likelihood ratio test increases. The plot also shows that the larger differences between p_0 and p_1 produces the better power curves. Simulations were also performed to verify the significance level of the test. The results were good and essentially the same as Table 1.

Example 4: A Real Data Example

Unfortunately no suitable unclassified data with categorical variables comparing nuclear explosions to earthquakes are available for this report. However, there is a considerable amount of mining explosion data available as training data. Therefore, to illustrate the method developed here, we have applied the bootstrap generalized likelihood ratio test to observations at the ARCESS seismic array in Norway which consist of mining blasts from two separate mines (HB6 and HD9) located in the Kola Peninsula of the former Soviet Union. (For other applications of the bootstrap generalized likelihood ratio test to seismic event identification, see Fisk and Gray (1993); Fisk et al., (1993).) Fifteen blasts were observed from mine HB6 and sixteen blasts were observed from mine HD9.

The variables used here are day-of-the-week (DOW), slowness (inverse group velocity measured in seconds/degree) of Pn (SLOW), and rectilinearity of Pn (RECT). Pn is typically the first prominent portion of the seismogram to arrive for signals observed at regional distances (<2000 km). These data are part of a data set established by Sereno and Patnaik (1992) as a testbed for seismic signal identification problems. Other features are also available in this data set, but most have many missing data values, a problem we are currently addressing.

A histogram plot of DOW is plotted in Figure 3 for the two sets of mining blasts. Note that the HD9 blasts occur predominantly on day 6, while the HB6 blasts occur more uniformly throughout the week. Dot plots of the continuous variables are shown in Figure 4. SLOW exhibits relatively good separation, while there is considerable overlap for RECT.

In order to assess the value of the discrete variable we considered cases in which DOW is either included or excluded. Since the day on which an event occurred has no influence on the seismogram, we treated the continuous variables as independent of DOW. Furthermore, we assumed unequal covariance matrices since the variances for SLOW are significantly different. Setting the significance level at 0.01 and 0.05, we estimated the power using the bootstrap with and without DOW. Table 3 gives the results using both continuous variables, while Table 4 gives the results using only RECT, with and without DOW. Since SLOW is such a strong discriminator, Table 4 better demonstrates the power that may be gained by making use of an available discrete feature.

Table 3. Bootstrap estimates of power using both SLOW and RECT.

Significance	DOW excluded	DOW included
0.01	0.962	0.982
0.05	0.980	0.986

Table 4. Bootstrap estimates of power using RECT.

Significance	DOW excluded	DOW included
0.01	0.266	0.377
0.05	0.529	0.736

The power was estimated in these tables using a parametric bootstrap approach. Specifically, given the training samples of size $N_0 = 15$ and $N_1 = 16$ available from the two mines, $\pi_0 = \text{HB6}$ and $\pi_1 = \text{HD9}$, ML estimates of the associated parameters are obtained. For these data, the bootstrap is used to estimate the α-level critical value by simulating $B = 499$ replications. Each replication consists of training samples of sizes

N_0 and N_1 from the models fit to π_0 and π_1 along with an observation to be classified which is generated according to the model for π_0. As in the previous examples, the α-level critical value, λ_α^*, was obtained from the likelihood ratio statistics calculated from these replicates. The power is then estimated by again simulating B bootstrap replications, where each replicate consists of training samples of sizes N_0 and N_1 from the models fit to π_0 and π_1 along with an observation to be classified which this time is generated according to the model for π_1. The power is estimated as the proportion of the resulting B likelihood ratio statistics that are less than or equal to λ_α^*. A cross-validation procedure was also considered, and it gave results similar to those shown here. Efron (1983) has suggested an alternative bootstrap approach to remove the bias from the cross-validation estimate.

4. Concluding Remarks

When one needs to classify an individual with one of the misclassification probabilities under control but does not know the exact or limiting distribution of the statistic for classification, the bootstrap likelihood ratio method is shown to be useful. The statistic used for classification is derived from the likelihood ratio, and its limiting distribution furnishing the discriminant cut-off point is approximated successfully by the parametric bootstrap.

The bootstrap likelihood ratio statistic is shown to compete well with the statistics W and Z whose limiting distributions are known, for moderate sample sizes when two multivariate normal distributions with equal covariance matrices are considered. It also performs quite well for both the multivariate normal case with unequal covariance matrices and the case of a mixture of binary and normal variates, where classical classification rules cannot control the probability of misclassification. Moreover, the methodology considered here can be applied to any non-normal discrete or continuous variable, and to any mixture of continuous and discrete variables, whenever the MLEs exist. It should be noted that the precision of the test depends on the sample sizes N_0 and N_1, and the bootstrap replication size B. Small sample sizes may result in MLEs for the parametric bootstrap which are not close to the true parameter values. Adequate sample sizes for different dimensions of the classification variable may need to be studied. Finally, it should be noted that the method applied here could be applied to any test of hypothesis based on the generalized likelihood ratio. Actually, the approach considered here of calculating λ based on normal likelihoods and finding λ^*_α, should be a sensible approach for continuous, unimodal distributions. The robustness of this procedure is the topic of current research.

Appendix: Formulas Related to Examples

Example 1

$\mu^{(i)}$ is estimated by $\bar{\mathbf{v}}^{(i)} = \sum_{j=1}^{N_i} \mathbf{v}_j^{(i)}/N_i$ and $\sum$ is estimated by

$$\mathrm{S} = \frac{(N_0 - 1)\mathrm{S}^{(0)} + (N_1 - 1)\mathrm{S}^{(1)}}{N_0 + N_1 - 2}\,, \tag{A1}$$

where $\mathrm{S}^{(i)} = \sum_{j=1}^{N_i}(\mathbf{v}_j^{(i)} - \bar{\mathbf{v}}^{(i)})\;(\mathbf{v}_j^{(i)} - \bar{\mathbf{v}}^{(i)})'/(N_i - 1)$, $i = 0, 1$. Anderson's W statistic is given by

$$W = \left[\mathbf{v} - \tfrac{1}{2}\left(\bar{\mathbf{v}}^{(0)} + \bar{\mathbf{v}}^{(1)}\right)\right]' \mathrm{S}^{-1}\left(\bar{\mathbf{v}}^{(0)} - \bar{\mathbf{v}}^{(1)}\right). \tag{A2}$$

Under the null hypothesis H_0, the MLEs of $\mu^{(0)}$, $\mu^{(1)}$, and Σ are

$$\hat{\mu}_0^{(0)} = (N_0\bar{\mathbf{v}}^{(0)} + \mathbf{v})/(N_0 + 1),$$

$$\hat{\mu}_0^{(1)} = \overline{\mathbf{v}}^{(1)},$$

$$\hat{\Sigma}_0 = \frac{1}{N_0 + N_1 + 1}\left[\mathrm{A} + \frac{N_0}{N_0 + 1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(0)}\right)\left(\mathbf{v} - \overline{\mathbf{v}}^{(0)}\right)'\right],$$

where $\mathbf{A} = \sum_{i=0}^{1} \sum_{j=1}^{N_i} (\mathbf{v}_j^{(i)} - \overline{\mathbf{v}}^{(i)})\ (\mathbf{v}_j^{(i)} - \overline{\mathbf{v}}^{(i)})' = (N_0 + N_1 - 2)\mathrm{S}$. Under the alternative hypothesis H_1, the MLEs of the parameters are

$$\hat{\mu}_1^{(0)} = \overline{\mathbf{v}}^{(0)}$$

$$\hat{\mu}_1^{(1)} = (N_1\overline{\mathbf{v}}^{(1)} + \mathbf{v})/(N_1 + 1),$$

$$\hat{\Sigma}_0 = \frac{1}{N_0 + N_1 + 1}\left[\mathrm{A} + \frac{N_1}{N_1 + 1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(1)}\right)\left(\mathbf{v} - \overline{\mathbf{v}}^{(1)}\right)'\right].$$

In this case the likelihood ratio given in (3), with $\hat{\theta}_0^{(0)} = (\hat{\mu}_0^{(0)}, \hat{\Sigma}_0)$, $\hat{\theta}_0^{(1)} = (\hat{\mu}_0^{(1)}, \hat{\Sigma}_0)$, $\hat{\theta}_1^{(0)} = (\hat{\mu}_1^{(0)}, \hat{\Sigma}_1)$, and $\hat{\theta}_1^{(1)} = (\hat{\mu}_1^{(1)}, \hat{\Sigma}_1)$ is, therefore,

$$\left[\frac{N + \frac{N_1}{N_1 + 1}(\mathbf{v} - \overline{\mathbf{v}}^{(1)})'\mathrm{S}^{-1}(\mathbf{v} - \overline{\mathbf{v}}^{(1)})}{N + \frac{N_0}{N_0 + 1}(\mathbf{v} - \overline{\mathbf{v}}^{(0)})'\mathrm{S}^{-1}(\mathbf{v} - \overline{\mathbf{v}}^{(0)})}\right]^{-(N_0 + N_1 + 1)/2}, \tag{A3}$$

where $N = N_0 + N_1 - 2$. The likelihood ratio (A3) is characterized by John's Z statistic,

$$Z = \frac{1}{2}\left[\frac{N_1}{N_1 + 1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(1)}\right)'\mathrm{S}^{-1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(1)}\right) - \frac{N_0}{N_0 + 1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(0)}\right)'\mathrm{S}^{-1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(0)}\right)\right].$$

Thus

$$\lambda = \log\left\{N + \frac{N_1}{N_1+1}\,(\mathbf{v} - \overline{\mathbf{v}}^{(1)})'\,\mathbf{S}^{-1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(1)}\right)\right\}$$

$$-\log\left\{N + \frac{N_0}{N_0+1}\,(\mathbf{v} - \overline{\mathbf{v}}^{(0)})'\,\mathbf{S}^{-1}(\mathbf{v} - \overline{\mathbf{v}}^{(0)})\right\}. \qquad \text{(A4)}$$

Example 2

The quadratic discriminant function is given by

$$Q = \frac{1}{2}\log\left(\frac{|\mathbf{S}^{(1)}|}{|\mathbf{S}^{(0)}|}\right) + \frac{1}{2}\,[(\mathbf{v} - \overline{\mathbf{v}}^{(1)})'(\mathbf{S}^{(1)})^{-1}(\mathbf{v} - \overline{\mathbf{v}}^{(1)})$$

$$-(\mathbf{v} - \overline{\mathbf{v}}^{(0)})'(\mathbf{S}^{(0)})^{-1}\,(\mathbf{v} - \overline{\mathbf{v}}^{(0)})\,]\,. \qquad \text{(A5)}$$

The MLEs of $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\mu}^{(1)}$, $\Sigma^{(0)}$, $\Sigma^{(1)}$ under H_0 are

$$\hat{\boldsymbol{\mu}}_0^{(0)} = (N_0\overline{\mathbf{v}}^{(0)} + \mathbf{v})/(N_0 + 1),$$

$$\hat{\boldsymbol{\mu}}_0^{(1)} = \overline{\mathbf{v}}^{(1)}\,,$$

$$\hat{\Sigma}_0^{(0)} = \frac{1}{N_0+1}\left[\mathbf{A}^{(0)} + \frac{N_0}{N_0+1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(0)}\right)\left(\mathbf{v} - \overline{\mathbf{v}}^{(0)}\right)'\right],$$

$$\hat{\Sigma}_0^{(1)} = \frac{1}{N_1}\,\mathbf{A}^{(1)}\,,$$

where $\mathbf{A}^{(i)} = \sum_{j=1}^{N_i}\,(\mathbf{v}_j^{(i)} - \overline{\mathbf{v}}^{(i)})\,(\mathbf{v}_j^{(i)} - \overline{\mathbf{v}}^{(i)})'$, $i = 0, 1$. Under the alternative hypothesis H_1, the MLEs are

$$\hat{\boldsymbol{\mu}}_1^{(0)} = \overline{\mathbf{v}}^{(0)},$$

$$\hat{\mu}_1^{(1)} = (N_1\overline{\mathbf{v}}^{(1)} + \mathbf{v})/(N_1 + 1),$$

$$\hat{\Sigma}_1^{(0)} = \frac{1}{N_0}\mathbf{A}^{(0)},$$

$$\hat{\Sigma}_1^{(1)} = \frac{1}{N_1 + 1}\left[\mathbf{A}^{(1)} + \frac{N_1}{N_1 + 1}\left(\mathbf{v} - \overline{\mathbf{v}}^{(1)}\right)\left(\mathbf{v} - \overline{\mathbf{v}}^{(1)}\right)'\right].$$

The log likelihood ratio statistic is given by

$$\lambda = \frac{1}{2}\log\left(\frac{|\mathbf{S}^{(1)}|}{|\mathbf{S}^{(0)}|}\right) + \frac{1}{2}\left[(N_1 + 1)\log\left\{(N_1 - 1) + \frac{N_1}{N_1+1}(\mathbf{v} - \overline{\mathbf{v}}^{(1)})'(\mathbf{S}^{(1)})^{-1}(\mathbf{v} - \overline{\mathbf{v}}^{(1)})\right\}\right.$$

$$\left. - (N_0 + 1)\log\left\{(N_0 - 1) + \frac{N_0}{N_0+1}(\mathbf{v} - \overline{\mathbf{v}}^{(0)})'(\mathbf{S}^{(0)})^{-1}(\mathbf{v} - \overline{\mathbf{v}}^{(0)})\right\}\right] + C(N_0, N_1) \qquad \text{(A6)}$$

where $\mathbf{S}^{(i)} = \mathbf{A}^{(i)}/(N_i - 1)$, $i = 0, 1$, and

$$C(N_0, N_1) = \log\left[\frac{(N_0-1)^{(N_0+1-p)/2}\ (N_0+1)^{(N_0+1)p/2}\ N_1^{N_1p/2}}{(N_1-1)^{(N_1+1-p)/2}\ N_0^{N_0p/2}\ (N_1+1)^{(N_1+1)p/2}}\right].$$

Example 3

We consider the log likelihood ratio statistic under the scenario discussed in *Example 3*, i.e. the new individual $\mathbf{v}$ to be classified has discrete components that place it into cell l. The likelihood functions on the numerator and denominator of (3) are given by

$$L_l^{(i)}L = \{(2\pi)^q\}^{-(N_0+N_1+1)/2}|\Sigma^{(0)}|^{-N_0/2}\ |\Sigma^{(1)}|^{-N_1/2}\ |\Sigma^{(i)}|^{-1/2}$$

$$\cdot\left\{\prod_{h=0}^{1}\prod_{m=1}^{k}(p_m^{(h)})^{n_m^{(h)}}\right\}\left(p_l^{(i)}\right)$$

$$\cdot\ exp\left[-\frac{1}{2}\left\{\sum_{h=0}^{1}\sum_{m=1}^{k}\left(\sum_{s=1}^{n_{\mathrm{m}}^{(h)}}(\mathbf{x}_{sm}^{(h)} - \boldsymbol{\mu}_m^{(h)})'(\Sigma^{(h)})^{-1}(\mathbf{x}_{sm}^{(h)} - \boldsymbol{\mu}_m^{(h)})\right)\right.\right.$$

$$\left.\left. + (\mathbf{x}-\boldsymbol{\mu}_l^{(i)})'(\Sigma^{(i)})^{-1}(\mathbf{x}-\boldsymbol{\mu}_l^{(i)})\right\}\right],\quad i = 0, 1.$$

Under H_0 the MLEs of $p_m^{(i)}$, $\boldsymbol{\mu}_m^{(i)}$, $\Sigma^{(i)}$ are

$$\hat{p}_{m0}^{(0)} = n_m^{(0)}/(N_0 + 1),\quad m = 1, \ldots, l-1, l+1, \ldots, k,$$

$$\hat{p}_{l0}^{(1)} = \left(n_l^{(0)} + 1\right)/(N_0 + 1),$$

$$\hat{\mu}_{m0}^{(0)} = \overline{\mathbf{x}}_m^{(0)},\quad m = 1, \ldots, l-1, l+1, \ldots, k,$$

$$\hat{\mu}_{l0}^{(0)} = (n_l^{(0)}\,\overline{\mathbf{x}}_l^{(0)} + \mathbf{x})/(n_l^{(0)} + 1),$$

$$\hat{\Sigma}_1^{(0)} = \frac{1}{N_0+1}\left[\mathbf{A}^{(0)} + \frac{N_l^{(0)}}{N_l^{(0)} + 1}(\mathbf{x}-\overline{\mathbf{x}}_l^{(0)})(\mathbf{x}-\overline{\mathbf{x}}_l^{(0)})'\right].$$

$$\hat{p}_{m0}^{(1)} = n_m^{(1)}/N_1,\quad m = 1, \ldots, k,$$

$$\hat{\mu}_{m0}^{(1)} = \overline{\mathbf{x}}_m^{(1)},\quad m = 1, \ldots, k,$$

$$\hat{\Sigma}_0^{(1)} = \frac{1}{N_1}\ \mathbf{A}^{(1)},$$

where $\overline{\mathbf{x}}_m^{(i)} = \sum_{s=1}^{n_m^{(i)}}\mathbf{x}_{sm}^{(i)}/n_m^{(i)}$, $\mathbf{A}_m^{(i)} = \sum_{s=1}^{n_m^{(i)}}(\mathbf{x}_{sm}^{(i)} - \overline{\mathbf{x}}_m^{(i)})(\mathbf{x}_{sm}^{(i)} - \overline{\mathbf{x}}_m^{(i)})'$, $m = 1, \ldots, k$, and $\mathbf{A}^{(i)} = \sum_{m=1}^{k}\mathbf{A}_m^{(i)}$. Under the alternative hypothesis H_1 the MLEs are

$$\hat{p}_{m1}^{(0)} = n_m^{(0)}/N_0, \ \ m = 1, \ldots, k,$$

$$\hat{\mu}_{m1}^{(0)} = \bar{\mathbf{x}}_m^{(0)}, \ \ m = 1, \ldots, k,$$

$$\hat{\Sigma}_1^{(0)} = \frac{1}{N_0} \ \mathbf{A}^{(0)} \ ,$$

$$\hat{p}_{m1}^{(1)} = n_m^{(1)}/(N_1 + 1), \ \ m = 1, \ldots, l-1, l+1, \ldots, k,$$

$$\hat{p}_{l1}^{(1)} = (n_l^{(1)} + 1)/(N_1 + 1),$$

$$\hat{\mu}_{m1}^{(1)} = \ \bar{\mathbf{x}}_m^{(1)}, m = 1, \ldots, l-1, \ l+1, \ldots, k,$$

$$\hat{\mu}_{l1}^{(1)} = (n_l^{(1)} \, \bar{\mathrm{x}}_l^{(1)} + \mathrm{x})/(n_l^{(1)} + 1) \ ,$$

$$\hat{\Sigma}_1^{(1)} = \frac{1}{N_1 + 1} \left[\mathbf{A}^{(1)} + \frac{N_l^{(1)}}{N_l^{(1)} + 1} \, (\mathbf{x} - \bar{\mathbf{x}}_l^{(1)})(\mathbf{x} - \bar{\mathbf{x}}_l^{(1)})' \right].$$

Since the exponential term of $L_l^{(i)}L$ after replacing the parameters by their MLEs, is $exp\{-(1/2)q(N_0 + N_1 + 1)\}$ for $i = 0, 1$, the log likelihood ratio statistic is given by

$$\lambda = log\left\{ \left\{ \prod_{i=0}^{1} \prod_{m=1}^{k} \left(\frac{\hat{p}_{m0}^{(i)}}{\hat{p}_{m1}^{(i)}}\right)^{n_m^{(i)}} \right\} \left(\frac{\hat{p}_{l0}^{(0)}}{\hat{p}_{l1}^{(1)}}\right) \left\{ \prod_{i=0}^{1} \left(\frac{|\hat{\Sigma}_0^{(\mathrm{i})}|}{|\hat{\Sigma}_1^{(\mathrm{i})}|}\right)^{-N_i/2} \right\} \left(\frac{|\hat{\Sigma}_0^{(0)}|}{|\hat{\Sigma}_1^{(1)}|}\right)^{-1/2} \right\}. \qquad \text{(A7)}$$

References

Anderson, T. W. (1951). "Classification by Multivariate Analysis," *Psychometrika*, 16, 631-650.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis,* New York: John Wiley & Sons.

Anderson, T. W. (1973), "An Asymptotic Expansion of the Distribution of the Standardized Classification Statistic *W*," *Annals of Statistics*, 1, 964-972.

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis 2nd Ed.*, John Wiley & Sons, New York.

Baumgardt, D.R., J. Carney, M. Maxson and S. Carter (1992), "Evaluation of Regional Discriminants using the Intelligent Seismic Event Identification System, Semi-Annual Technical Report SAS-TR-93-38, ENSCO, Inc., Springfield, VA.

Bickel, P. J. and Freedman, D. A. (1981), "Some Asymptotic Theory for the Bootstrap", *Annals of Statistics,* 6, 1196-1217.

Efron, B., (1979), "Bootstrap Methods: Another Look at the Jackknife,"*Annals of Statistics,* 7, 1-26.

Efron, B., (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia.

Efron, B., (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, 78, 316-331.

Fisk, M.D. and H.L. Gray (1993), "Event Identification Analysis of the Novaya Zemlya Event on 31 December 1992 Using Outlier and Classification Likelihood Ratio Tests," MRC-R-1449, Mission Research Corp., Santa Barbara, CA.

Fisk, M.D. H.L. Gray, and G.D. McCartor (1993), "Applications of Generalized Likelihood Ratio Tests to Seismic Event Identification," PL-TR-93-2221, Phillips Laboratory, Hanscom AFB, MA, ADA279479.

John, S., (1960), "On Some Classification Problems, " *Sankhyá*, Series A 22, 301-308.

John, S., (1963), "On Classification by the Statistics R and Z," *Annals of the Institute of Statistical Mathematics*, 14, 237-246.

Kanazawa, M., (1979), "The Asymptotic Cut-off Point and Comparison of Error Probabilities in Covariate Discriminant Analysis," *Journal of the Japan Statistical Society,* 9, 7-17.

Knoke, J. D., (1982), "Discriminant Analysis with Discrete and Continuous Variables," *Biometrics*, 38, 191-200.

Krzanowski, W. J. ,(1975), "Discrimination and Classification using Both Binary and Continuous Variables," *Journal of the American Statistical Association*, 70, 782-790.

Krzanowski, W. J., (1979), "Some Linear Transformations for Mixtures of Binary and Continuous Variables, with Particular Reference to Linear Discriminant Analysis", *Biometrika,* 66, 33-39.

Krzanowski, W. J., (1980), "Mixtures of Continuous and Categorical Variables in Discriminant Analysis," *Biometrics,* 36, 493-499.

Krzanowski, W. J., (1982), "Mixtures of Continuous and Categorical Variables in Discriminant Analysis: A Hypothesis-Testing Approach," *Biometrics*, 38, 991-1002.

McLachlan, G. J., (1987), "On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture," *Applied Statistics,* 36 (3), 318-324.

Olkin, I. and Tate, R. F., (1961) "Multivariate Correlation Models with Mixed Discrete and Continuous Variables," *Annals of Mathematical Statistics,* 32, 448-65.

Seber, G. A. F., (1984), *Multivariate Observations*, New York: John Wiley & Sons.

Sereno, T.J. and G.B. Patnaik, (1992), "Data to Test and Evaluate the Performance of Neural Network Architectures for Seismic Signal Discrimination," PL-TR-92-2110(I), Phillips Laboratory, Hanscom AFB, MA, ADA254413.

Sereno, T.J. and D. Wahl, (1993), "A Fuzzy-logic Approach to Regional Seismic Event Identification: Application to the Novaya Zemlya Event on 31 December 1992," SAIC-93/1156, Science Applications International Corp., San Diego, CA.

Shumway, R.H. (1988), *Applied Statistical Time Series Analysis*, Englewood Cliffs, New Jersey: Prentice Hall.

Tu, C. and Han, C., (1982), "Discriminant Analysis Based on Binary and Continuous Variables," *Journal of the American Statistical Association*, 77, 447-454.

Welch, B. L., (1939), "Note on Discriminant Functions," *Biometrika*, 31, 218-220.

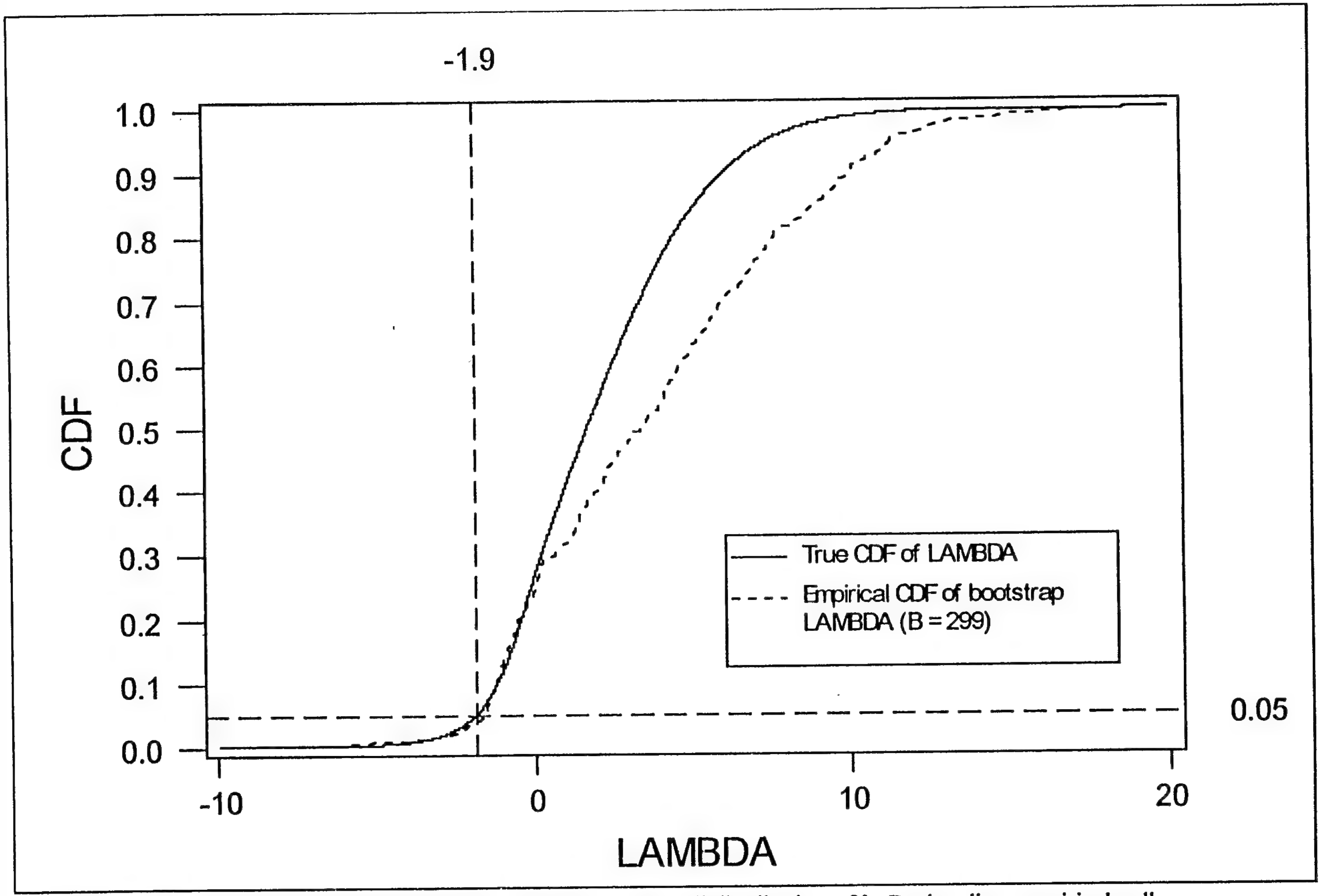


Figure 1(a). Plots of distribution functions. Solid line: true null distribution of λ. Broken line: empirical null distribution of λ. $N_0 = 10$, $N_1 = 15$.

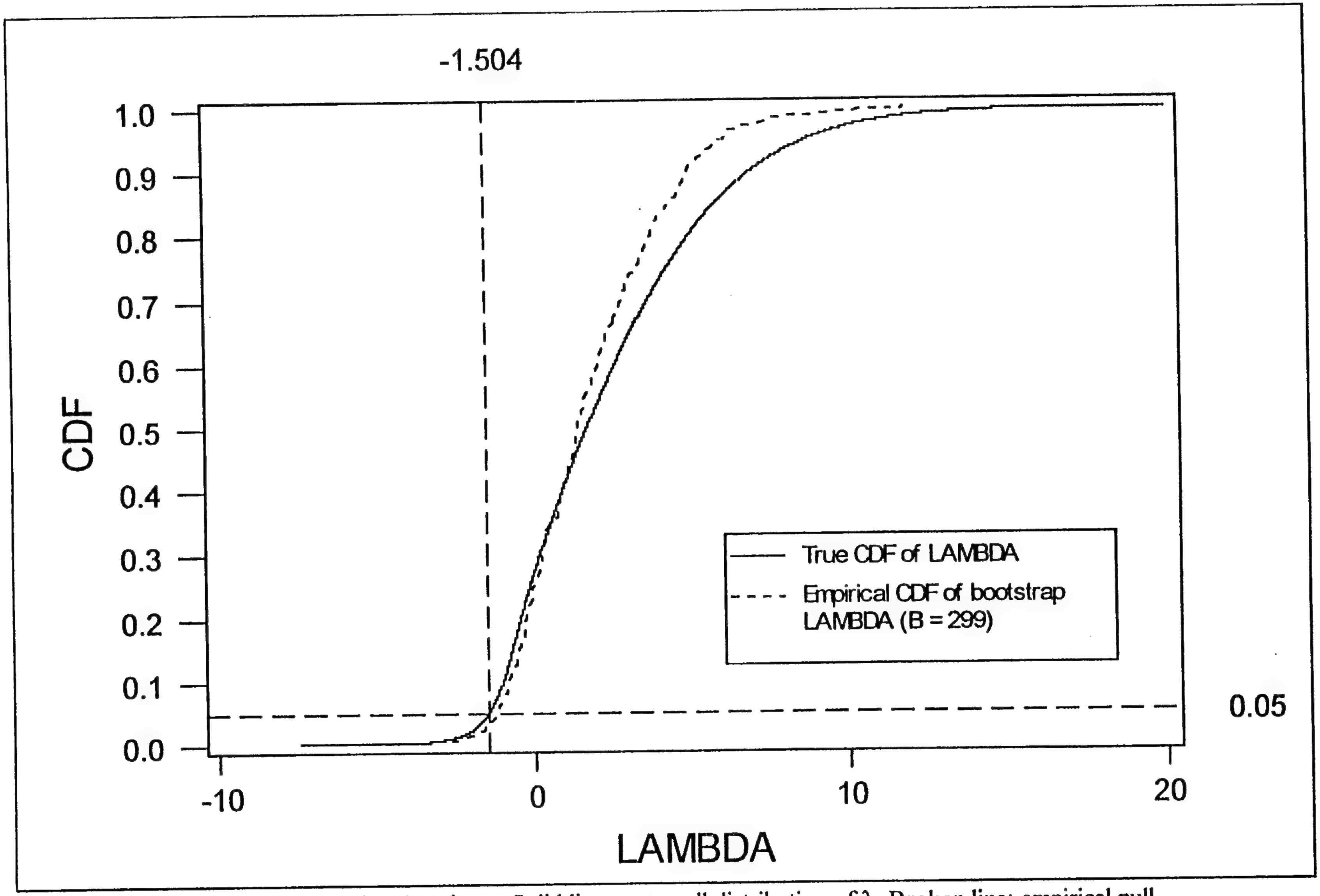


Figure 1(b). Plots of distribution functions. Solid line: true null distribution of λ. Broken line: empirical null distribution of λ. $N_0 = 30$, $N_1 = 45$.

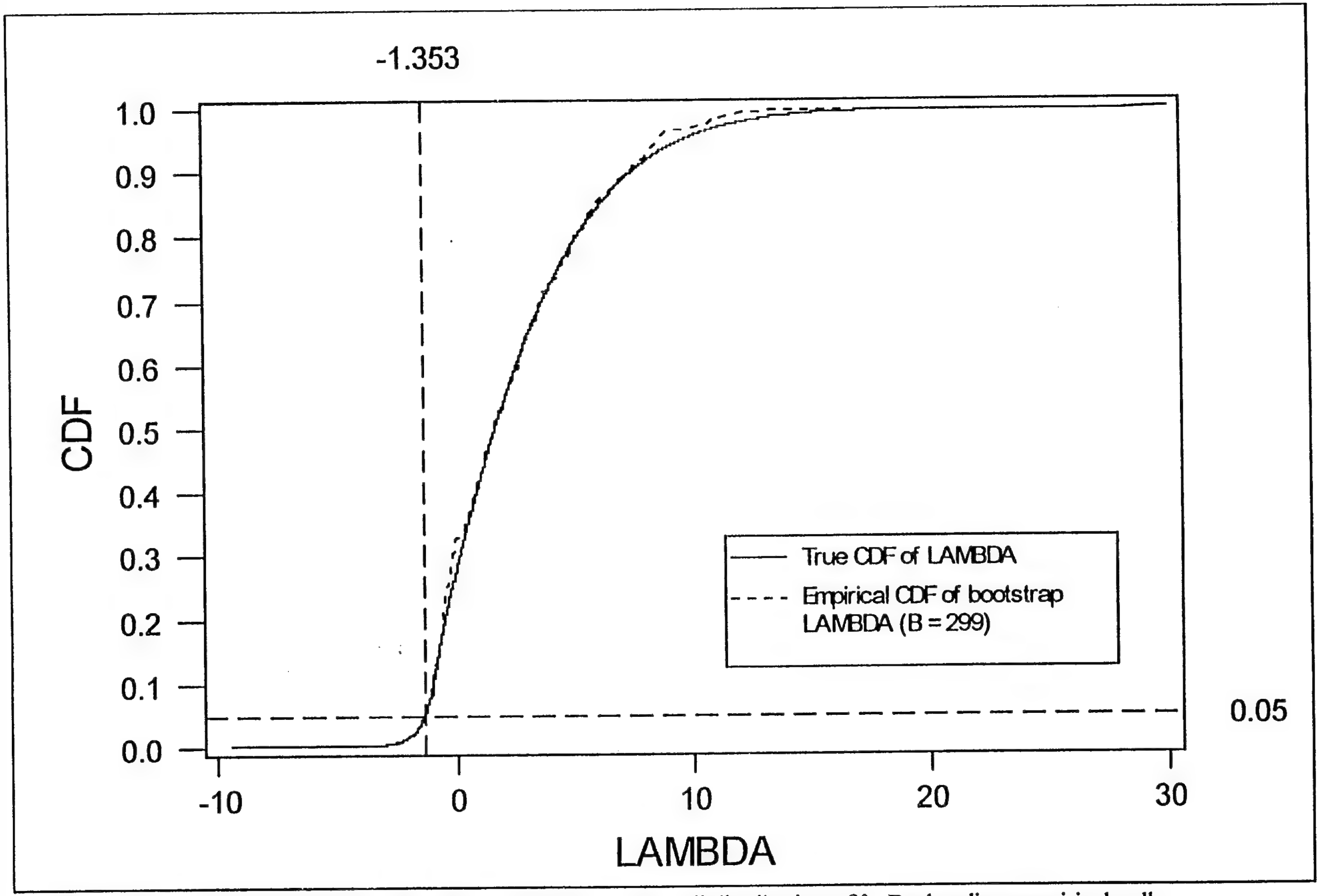


Figure 1(c). Plots of distribution functions. Solid line: true null distribution of λ. Broken line: empirical null distribution of λ. $N_0 = 100$, $N_1 = 150$.

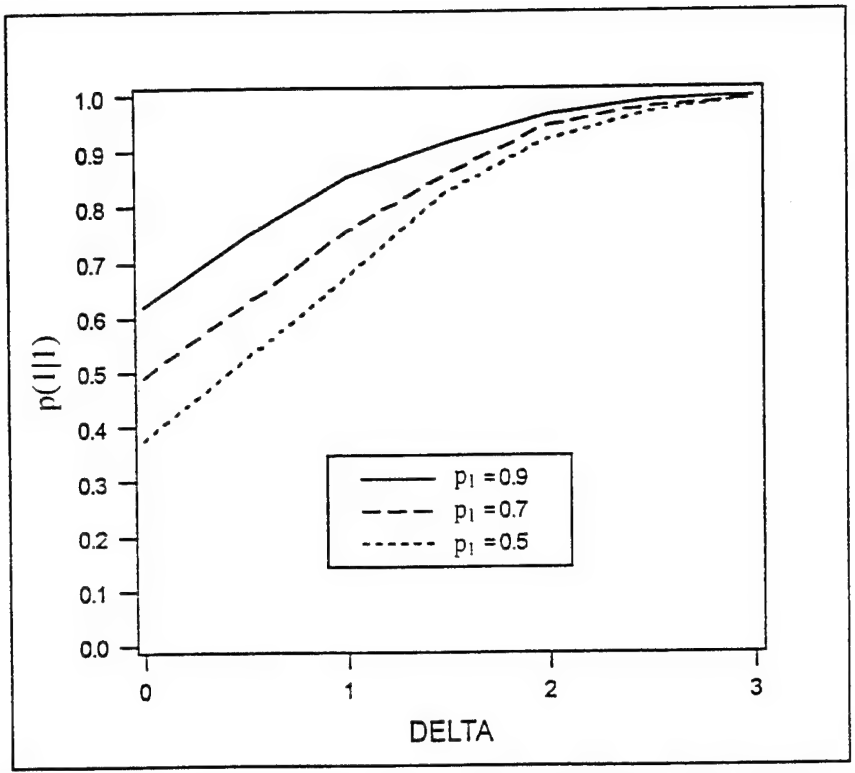


Figure 2. Power curves of bootstrap λ with mixed binary and continuous variables. $p_0 = 0.1$. DELTA denotes Δ.

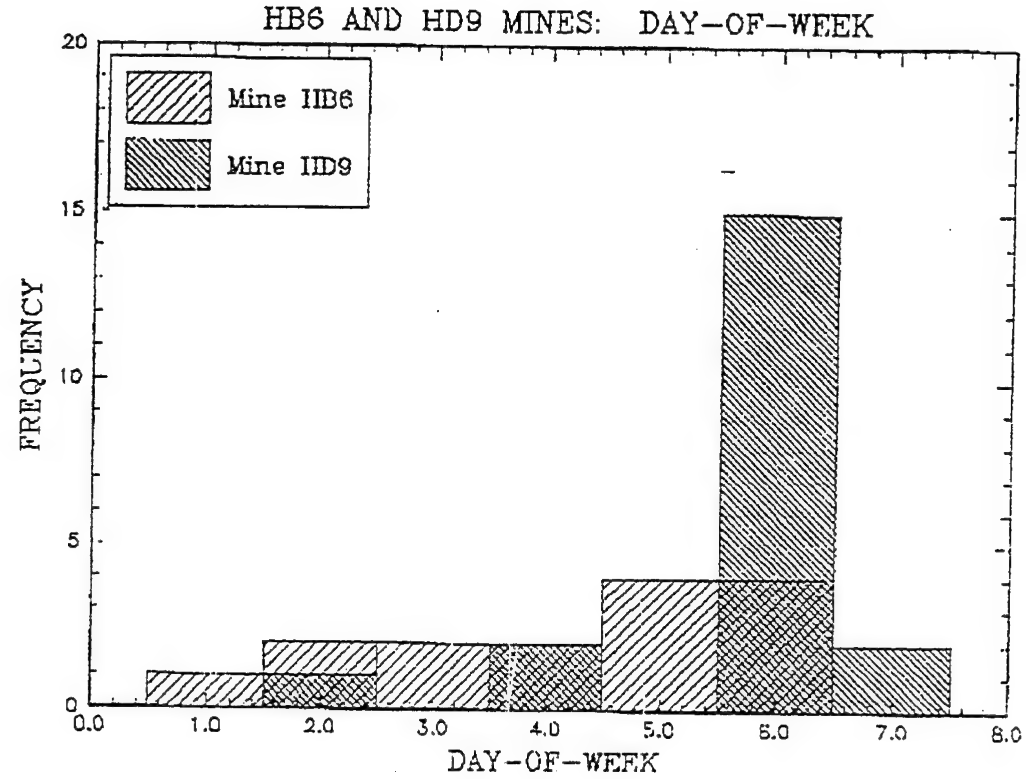


Figure 3. Histogram plot of the categorical variable day-of-week.

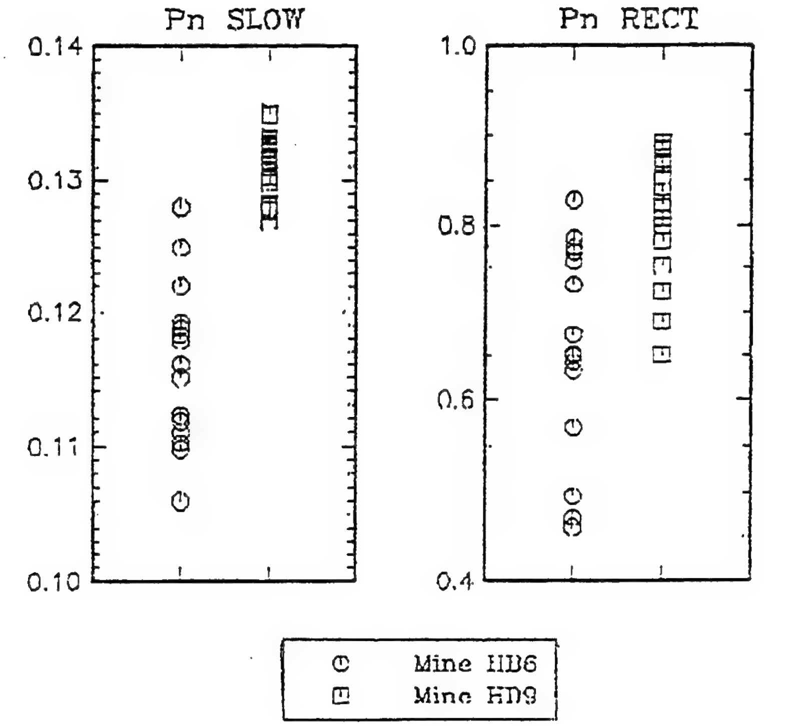


Figure 4. Dot plots of continuous variables used to classify mining blasts.

A HYPOTHESIS-TESTING APPROACH TO DISCRIMINANT ANALYSIS WITH MIXED CATEGORICAL AND CONTINUOUS VARIABLES WHEN DATA ARE MISSING

by

James W. Miller, Wayne A.Woodward, Henry L. Gray,
Mark A. Fisk*, Gary D. McCartor

Southern Methodist University
Dallas, Texas 75275

*Mission Research Corporation
Santa Barbara, California 93102

Technical Report No. SMU/DS/TR-273

JULY 1994

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

Sponsored by
Advanced Research Projects Agency (DoD)
Nuclear Monitoring Research Office
Issued by the Phillips Laboratory under Contract No. F19628-93-C-0199

A HYPOTHESIS-TESTING APPROACH TO DISCRIMINANT ANALYSIS WITH MIXED CATEGORICAL AND CONTINUOUS VARIABLES WHEN DATA ARE MISSING

James W. Miller, Wayne A. Woodward, and Henry L. Gray

ABSTRACT

In this paper we consider the problem of discriminant analysis with discrete (categorical) and continuous variables with data missing at random. We use a hypothesis-testing approach based on the generalized likelihood ratio as proposed by Baek, et al. (1994). We use bootstrapping to determine critical values in order to control the Type I error rate. We present three algorithms for dealing with this case, each assuming a different model for the data: (1) The INDICATOR algorithm replaces categorical variables with indicator variables, and treats these as if they were continuous; (2) the FULL algorithm assumes a multinomial distribution for the discrete part, and a multivariate normal distribution (with mean and covariances depending on the discrete part) as the conditional distribution of the continuous part given the discrete part; and (3) the COMMON algorithm assumes a multinomial distribution for the discrete part, and a multivariate normal distribution (with only the means depending on the discrete part) as the conditional distribution of the continuous part given the discrete part. (That is, a common covariance matrix is assumed across all multinomial cells.) The performance of these algorithms is compared through a simulation study. While the INDICATOR algorithm seems to have highest power, it also tends to display a higher Type I error rate than desired. The FULL and the COMMON algorithms have very similar power, but the COMMON algorithm appears to control the Type I error rate most effectively, and is least susceptible to problems occurring when some multinomial cells are sparsely represented.

1. Introduction

In Baek, Gray, Woodward, Miller, and Fisk (1994) (subsequently abbreviated BGWMF) techniques are given for a hypothesis-testing approach to discriminant analysis in which one wishes to control one of the probabilities of misclassification. Methods are presented for continuous variables only, as well as for a mixture of continuous and categorical variables. Essentially, the hypothesis-testing approach based on the ratio of maximized likelihood functions proposed by Krzanowski (1982) is employed and the test statistic is bootstrapped in order to estimate critical values for the allocation rule in such a way that the error rate is controlled. In Miller, Gray, and Woodward (1993) (subsequently abbreviated (MGW)), a similar hypothesis-testing approach is used for discriminant analysis and outlier detection in the presence of missing data. The EM algorithm (Dempster, Laird, and Rubin (1977)) is employed to obtain maximum likelihood estimates of model parameters and compute the maximized likelihoods based on the available data. That paper, however, only considers the case in which all variables are continuous and, in fact, normally distributed.

In this report, we wish to consider the remaining case in which we have a mixture of continuous and categorical variables used as discriminants, and also missing data, potentially in both the training sets and in the new observation to be classified. Once again, we use a hypothesis-testing approach to classification and bootstrap the test statistic in order to control the probability of a particular type of misclassification. We present three algorithms for handling this situation:

(1) The INDICATOR algorithm - This algorithm begins by converting each categorical variable with j categories into j - 1 indicator variables. This results is a larger number of variables (unless all categorical variables are already binary, in which case the data set is unchanged). These indicator variables can be analyzed using techniques for quantitative data. In this algorithm we make the (obviously

incorrect) assumption that all variates are continuous and, in fact, normally distributed. We then perform discriminant analysis using the transformed data and the techniques of MGW.

(2) The FULL algorithm - Next, we model the joint distribution of each observation in the following manner: Suppose each observation consists of p categorical variables and q continuous variables. The categorical variables define r cells of a contingency table in which the observation could fall, where r is the product of the number of categories possible within each categorical variable. We assume that the observation will fall into cell i ($i = 1, ..., r$) with probability p_i, and that the conditional distribution of the continuous part given that the discrete part places the observation into cell i is multivariate normal with mean μ_i and Σ_i. We then employ the EM algorithm to obtain maximum likelihood estimates of parameters in this model and compute maximized likelihoods of the available data, and bootstrap the ratio of maximized likelihoods, as was done in BGWMF.

(3) The COMMON algorithm - This algorithm is essentially the same as the FULL algorithm, except that we assume a common covariance matrix across all multinomial cells. That is, the conditional distribution of the continuous part given that the discrete part places the observation into cell i is assumed multivariate normal with mean μ_i and Σ, with Σ no longer depending on i. This reduces considerably the number of parameters that need to be estimated and makes possible calculation of the likelihood ratio statistic when some cells may be sparsely represented, or not represented at all.

Simulation studies are conducted to compare and contrast the performance of each of these procedures with regard to their ability to accurately control the Type I error rate, and with regard to their power.

2. Notation and Overview of the Generalized Likelihood Ratio Test Procedure

Suppose we wish to classify a (p+q)-dimensional random vector **V** into one of two populations π_1 or π_2. Suppose further that **V** can be partitioned as **V** = (**X**, **Y**), where **X** = $(X_1, X_2, ..., X_p)$ is a p-dimensional vector of categorical variables and **Y** = $(Y_1, Y_2, ..., Y_q)$ is a q-dimensional vector of continuous variables. Suppose that for i = 1, ..., p, the variable X_i takes on one of the r_i possible values 1, 2, ..., r_i. Then the vector **X** takes on one of $r = \Pi_{i=1}^{p} r_i$ possible values. We let Ψ denote the set of all possible values of the vector **X**. Finally, suppose that training samples $\{\mathbf{V}_i^{(1)}\}$, i = 1, ..., N_1 from π_1 and $\{\mathbf{V}_i^{(2)}\}$, i = 1, ..., N_2 from π_2, each having the same structure as **V**, are available, and that data may be missing at random from any part of **V** or from the training samples.

The generalized likelihood ratio test (GLRT) procedure for classifying **V** into π_1 or π_2 is based on a hypothesis testing approach. That is, the classification of **V** is done by testing

$$\begin{gathered} H_0: \mathbf{V}, \mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \dots, \mathbf{V}_{N_1}^{(1)} \in \pi_1; \mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \dots, \mathbf{V}_{N_2}^{(2)} \in \pi_2 \\ \text{versus} \\ H_1: \mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \dots, \mathbf{V}_{N_1}^{(1)} \in \pi_1; \mathbf{V}, \mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \dots, \mathbf{V}_{N_2}^{(2)} \in \pi_2. \end{gathered} \tag{1}$$

The two misclassification probabilities that we will be interested in are P(2|1) and P(1|2), where P(i|j) denotes the probability of classifying **V** into π_i when in fact $\mathbf{V} \in \pi_j$. We will refer to α = P(2|1) as the significance level for the test and P(2|2) as the power.

Let m denote the number of elements in **V** that are missing and let $\mathbf{V}_{(2)} = (\mathbf{X}_{(2)}, \mathbf{Y}_{(2)})$denote the (p - m)-variate vector of available data in **V**. Similarly, let $m_i^{(j)}$ denote the number of elements missing from $\mathbf{V}_i^{(j)}$ and let $\mathbf{V}_{i(2)}^{(j)}$ denote the (p - $m_i^{(j)}$)-variate vector of available data in $\mathbf{V}_i^{(j)}$ (j = 1, 2; i = 1, 2, ..., N_j). We assume that π_1 has joint density function $f(\mathbf{V}|\theta^{(1)})$ and that π_2 has joint density function $f(\mathbf{V}|\theta^{(2)})$, where f is some parametric density function with parameters $\theta^{(1)}$ and $\theta^{(2)}$ for populations π_1 and π_2, respectively. Then, under H_0, the likelihood of **V** and the training samples is given by

$$L_{01}(\theta^{(1)}|\, \mathbf{V}, \mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \dots, \mathbf{V}_{N_1}^{(1)}) L_{02}(\theta^{(2)}|\, \mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \dots, \mathbf{V}_{N_2}^{(2)}),$$

where (2)

$$L_{01}(\theta^{(1)}|\, \mathbf{V}, \mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \dots, \mathbf{V}_{N_1}^{(1)}) = f_2(\mathbf{V}|\theta^{(1)}) \prod_{i=1}^{N_1} f_{1i}(\mathbf{V}_i^{(1)}|\theta^{(1)}),$$

$$L_{02}(\theta^{(2)}|\, \mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \dots, \mathbf{V}_{N_2}^{(2)}) = \prod_{i=1}^{N_2} f_{2i}(\mathbf{V}_i^{(2)}|\theta^{(2)}),$$

$f_2(\mathbf{V}|\theta^{(1)})$ is the marginal density function for $\mathbf{V}_{(2)}$ evaluated at $\mathbf{V}_{(2)}$ with parameters $\theta^{(1)}$, and $f_{ji}(\mathbf{V}_i^{(j)}|\theta^{(j)})$ is the marginal density function for $\mathbf{V}_{i(2)}^{(j)}$ evaluated at $\mathbf{V}_{i(2)}^{(j)}$ with parameters $\theta^{(j)}$. Under H_1, the likelihood of **V** and the training samples is given by

$$L_{11}(\theta^{(1)}|\, \mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \dots, \mathbf{V}_{N_1}^{(1)}) L_{12}(\theta^{(2)}|\, \mathbf{V}, \mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \dots, \mathbf{V}_{N_2}^{(2)}),$$

where (3)

$$L_{11}(\theta^{(1)}|\, \mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \dots, \mathbf{V}_{N_1}^{(1)}) = \prod_{i=1}^{N_1} f_{1i}(\mathbf{V}_i^{(1)}|\theta^{(1)}),$$

$$L_{02}(\theta^{(2)}|\, \mathbf{V}, \mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \dots, \mathbf{V}_{N_2}^{(2)}) = f_2(\mathbf{V}|\theta^{(2)}) \prod_{i=1}^{N_2} f_{2i}(\mathbf{V}_i^{(2)}|\theta^{(2)}),$$

and $f_2(\mathbf{V}|\theta^{(2)})$ is the marginal density function for $\mathbf{V}_{(2)}$ evaluated at $\mathbf{V}_{(2)}$ with parameters $\theta^{(2)}$. We emphasize that these are the likelihood functions for the available data rather than the likelihood functions for the complete data since f_2 and f_{ji} ($j = 1, 2$; $i = 1, 2, \dots, N_j$) are marginal densities for the available part of each observation, rather than the likelihood functions for the complete data.

The GLRT procedure is based on the ratio

$$\mathrm{LR} = \frac{\sup\limits_{(\theta^{(1)},\,\theta^{(2)})} L_{01}(\theta^{(1)} |\, \mathbf{V}, \mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \ldots, \mathbf{V}_{N_1}^{(1)}) L_{02}(\theta^{(2)} |\, \mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \ldots, \mathbf{V}_{N_2}^{(2)})}{\sup\limits_{(\theta^{(1)},\,\theta^{(2)})} L_{11}(\theta^{(1)} |\, \mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \ldots, \mathbf{V}_{N_1}^{(1)}) L_{12}(\theta^{(2)} |\, \mathbf{V}, \mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \ldots, \mathbf{V}_{N_2}^{(2)})} \tag{4}$$

$$= \frac{L_{01}(\hat{\theta}_0^{(1)} |\, \mathbf{V}, \mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \ldots, \mathbf{V}_{N_1}^{(1)}) L_{02}(\hat{\theta}_0^{(2)} |\, \mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \ldots, \mathbf{V}_{N_2}^{(2)})}{L_{11}(\hat{\theta}_1^{(1)} |\, \mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \ldots, \mathbf{V}_{N_1}^{(1)}) L_{12}(\hat{\theta}_1^{(2)} |\, \mathbf{V}, \mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \ldots, \mathbf{V}_{N_2}^{(2)})},$$

where $\hat{\theta}_0^{(j)}$ and $\hat{\theta}_1^{(j)}$ are maximum likelihood estimates of $\theta^{(j)}$ ($j = 1, 2$) under the null and alternative hypotheses, respectively. That is, $\hat{\theta}_0^{(1)}$ is the MLE of $\theta^{(1)}$ based on the sample $\{\mathbf{V}, \mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \ldots, \mathbf{V}_{N_1}^{(1)}\}$, $\hat{\theta}_0^{(2)}$ is the MLE of $\theta^{(2)}$ based on the sample $\{\mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \ldots, \mathbf{V}_{N_2}^{(2)}\}$, and $\hat{\theta}_1^{(1)}$ is the MLE of $\theta^{(1)}$ based on the sample $\{\mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \ldots, \mathbf{V}_{N_1}^{(1)}\}$, and $\hat{\theta}_1^{(2)}$ is the MLE of $\theta^{(2)}$ based on the sample $\{\mathbf{V}, \mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \ldots, \mathbf{V}_{N_2}^{(2)}\}$.

Equivalently, the test procedure may be based on the statistic

$$\lambda = \log(\mathrm{LR}) = \lambda_{01} + \lambda_{02} - \lambda_{11} - \lambda_{12}, \tag{5}$$

where

$$\lambda_{01} = \log f_2(\mathbf{V}|\hat{\theta}_0^{(1)}) + \sum_{i=1}^{N_1} \log f_{1i}(\mathbf{V}_i^{(1)}|\hat{\theta}_0^{(1)}), \tag{6}$$

$$\lambda_{02} = \sum_{i=1}^{N_2} \log f_{2i}(\mathbf{V}_i^{(2)}|\hat{\theta}_0^{(2)}),$$

$$\lambda_{11} = \sum_{i=1}^{N_1} \log f_{1i}(\mathbf{V}_i^{(1)}|\hat{\theta}_1^{(1)}), \text{ and}$$

$$\lambda_{12} = \log f_2(\mathbf{V}|\hat{\theta}_1^{(2)}) + \sum_{i=1}^{N_1} \log f_{2i}(\mathbf{V}_i^{(2)}|\hat{\theta}_1^{(2)}).$$

A key step in evaluating λ for a given sample is the computation of the maximum likelihood estimates and the corresponding maximized log-likelihood functions λ_{01}, λ_{02}, λ_{11}, and λ_{12} in Equation (6). This is no trouble when the data are complete, as illustrated by (BGWMF). However, in the presence of missing data, the usual expressions for maximum likelihood estimates are no longer valid. In this case, maximum likelihood estimates are obtained via the EM algorithm (Dempster, Laird, and Rubin (1977)). The EM algorithm is an iterative procedure for obtaining parameter estimates which maximize the likelihood function of the available data. It involves two key steps:

(E -step) - Using current estimates $\hat{\theta}^{(k)}$ (where k now denotes the current iteration step, rather than designating π_1 or π_2), estimate the values of the complete data sufficient statistics by computing their expectations given the available data.

(M-step) - Determine the values of the parameters which maximize the likelihood for the complete data based on the current estimates of the complete data sufficient statistics, thus yielding $\hat{\theta}^{(k+1)}$.

The EM algorithm iteratively performs E- and M-steps until the sequence $\{\hat{\theta}^{(k)}\}$ converges to an adequate approximation to the MLE. To evaluate the test statistic λ of Equation (5), we must implement the EM algorithm four times. That is, $\hat{\theta}_0^{(1)}$ and λ_{01} are based on the sample $\{\mathbf{V}, \mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \dots, \mathbf{V}_{N_1}^{(1)}\}$, $\hat{\theta}_0^{(2)}$ and λ_{02} are based on the sample $\{\mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \dots, \mathbf{V}_{N_2}^{(2)}\}$, $\hat{\theta}_1^{(1)}$ and λ_{11} are based on the sample $\{\mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \dots, \mathbf{V}_{N_1}^{(1)}\}$, and $\hat{\theta}_1^{(2)}$ and λ_{12} are based on the sample $\{\mathbf{V}, \mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \dots, \mathbf{V}_{N_2}^{(2)}\}$.

The decision rule is described as follows: small values of λ provide evidence in favor of H_1, hence, $\mathbf{V}$ is classified into π_2 if $\lambda \leq \lambda_\alpha$, otherwise $\mathbf{v}$ is classified into π_1. The cut-off value λ_α is chosen so that $P(2|1) = \alpha$, the desired significance level for the test. Since the null distribution of λ is not known, the critical value is approximated by the parametric bootstrap (Efron 1979). For some large integer B, B bootstrap samples $\{\mathbf{V}^*, \mathbf{V}_1^{*(1)}, \mathbf{V}_2^{*(1)}, \dots, \mathbf{V}_{N_1}^{*(1)}\}$ are simulated from a distribution with density $f(\mathbf{V}|\hat{\theta}^{(1)})$ and B bootstrap samples $\{\mathbf{V}_1^{*(2)}, \mathbf{V}_2^{*(2)}, \dots, \mathbf{V}_{N_2}^{*(2)}\}$ are simulated from a distribution with density $f(\mathbf{V}|\hat{\theta}^{(2)})$, where $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ are MLEs obtained from the samples $\{\mathbf{V}_1^{(1)}, \mathbf{V}_2^{(1)}, \dots, \mathbf{V}_{N_1}^{(1)}\}$, and $\{\mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)}, \ldots, \mathbf{V}_{N_2}^{(2)}\}$, respectively. (Notice that in this case, $\hat{\theta}^{(1)} = \hat{\theta}_1^{(1)}$ and $\hat{\theta}^{(2)} = \hat{\theta}_0^{(2)}$.) When there are missing values, the simulated bootstrap samples should also have missing values in a configuration similar to that in the actual data. For each bootstrap sample, the test statistic λ is computed, thus generating a random sample $\{\lambda_1^*, \lambda_2^*, \ldots, \lambda_B^*\}$ of variates that have approximately the same distribution as λ under H_0. For an α-level test, the cut-off value λ_α^* is chosen as the α-th empirical quantile of $\{\lambda_1^*, \lambda_2^*, \ldots, \lambda_B^*\}$. Finally, $\mathbf{V}$ is classified into π_1 if $\lambda > \lambda_\alpha^*$; $\mathbf{V}$ is classified into π_2 if $\lambda \leq \lambda_\alpha^*$.

As was pointed out in (MGW), this test procedure is only an approximation to the true GLRT procedure since the critical value is obtained via bootstrapping and we may further relax our approximation to the true GLRT procedure by relaxing the number of iterations performed by the EM algorithm. That is, we may choose a stopping criterion for the EM algorithm that does not continue iteration until convergence has been obtained to a high degree of accuracy. Whatever the stopping criterion, bootstrapping the test statistic insures an approximate α-level test. As in (MGW), it would appear that very little power is lost by only performing a very few iterations of the EM algorithm, as opposed to carrying out iterations until MLEs are obtained with a high degree of accuracy. In Section 6, we often use only three iterations as standard practice in our simulation studies.

Our implementation of the GLRT procedure for discriminant analysis is summarized in Figure 1. Figures 2, 3, and 4 further describe the bootstrapping module, the computation of the test statistic λ, and the EM algorithm for obtaining MLEs. Each of the various algorithms discussed in this paper share this common skeletal structure. The differences lie in the type of model being assumed for the data, the corresponding implementation of the EM algorithm for obtaining MLEs, and the precise formulas used to evaluate the maximized log-likelihood functions.

3. The INDICATOR Algorithm

In our first attempt to implement an algorithm for discriminant analysis for mixed categorical and continuous variables with missing data, we desired to use the methods presented in MGW with as little adaptation as possible. One way to do this would be to treat the categorical variables as if they were continuous and use the procedures of MGW without any alteration at all. This is perhaps not such a bad idea if categorical variables have a large number of categories, if these categories have a natural ordering, and if the distribution of this variable has a somewhat normal shape. In most cases, however, these conditions are not satisfied and the procedure would be totally inappropriate.

A modification to the above approach is to replace each categorical variable with indicator variables in the following manner: We replace each categorical variable X_i $(i = 1, \dots, p)$ from $\mathbf{V}$ with the $r_i - 1$ indicator variables

$$W_{ij} = \mathbf{I}(X_i = j) = \begin{cases} 1 \text{ if } X_i = j \\ 0 \text{ otherwise} \end{cases} \qquad (j = 1, 2, \dots, r_i - 1). \tag{7}$$

Hence, the vector $\mathbf{X}$ of categorical variables gets replaced by a vector $\mathbf{W}$ of binary variables of length $\Sigma_{i=1}^{p} r_i - p$, producing the transformed vector $\tilde{\mathbf{V}} = (\mathbf{W}, \mathbf{Y})$. If X_i is missing in $\mathbf{X}$, then each W_{ij} $(j = 1, 2, \dots, r_i - 1)$ is missing in $\mathbf{W}$. We transform the training samples $\mathbf{V}_i^{(j)}$ $(j = 1, 2; i = 1, 2, \dots, N_j)$ in a similar manner producing $\tilde{\mathbf{V}}_i^{(j)}$ $(j = 1, 2; i = 1, 2, \dots, N_j)$.

Now, having transformed each observation, we classify $\mathbf{V}$ by classifying $\tilde{\mathbf{V}}$ according to the GLRT procedure as outlined in (MGW) for the continuous-variables-only case with missing data using the transformed data $\tilde{\mathbf{V}}$ and $\tilde{\mathbf{V}}_i^{(j)}$ $(j = 1, 2; i = 1, 2, \dots, N_j)$. That is, we proceed as if $\tilde{\mathbf{V}}$ and $\tilde{\mathbf{V}}_i^{(j)}$ $(j = 1, 2; i = 1, 2, \dots, N_j)$ were normally distributed, ignoring the fact that many of the components are binary. In simulation studies (see Section 6, below), we see that this method actually performs about as well as methods based on a more plausible model for the categorical variables, and is much easier to implement.

4. The FULL Algorithm

Next, we derive the GLRT procedure using a more plausible model for the distribution of **V**. In this case, we assume that the distribution of **X** follows a multinomial distribution in the sense that $\Pr[\mathbf{X} = \mathbf{x}] = p_{\mathbf{x}}$ for each $\mathbf{x} \in \Psi$, and the conditional distribution of **Y** given **X** = **x** is multivariate normal with mean $\mu_{\mathbf{x}}$ and covariance matrix and $\Sigma_{\mathbf{x}}$. Hence, $\theta = \{p_{\mathbf{x}}, \mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}; \mathbf{x} \in \Psi\}$ and

$$f(\mathbf{v}|\theta) = p_{\mathbf{x}}\mathrm{MVN}(\mathbf{y}|\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}), \tag{8}$$

where $\mathrm{MVN}(\mathbf{y}|\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$ denotes the value of the multivariate normal density function with parameters $\mu_{\mathbf{x}}$ and $\Sigma_{\mathbf{x}}$ evaluated at **y**.

The first step in deriving the GLRT procedure is to develop the EM algorithm for obtaining MLEs of θ given a collection of observations $(\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n)$ with missing values from such a population. The vectors $\mathbf{V}_i$ may be partitioned as $(\mathbf{X}_i, \mathbf{Y}_i)$, where $\mathbf{X}_i$ and $\mathbf{Y}_i$ are the vectors of categorical variables, and continuous variables respectively, and further partitioned as $(\mathbf{X}_{1i}, \mathbf{X}_{2i}, \mathbf{Y}_{1i}, \mathbf{Y}_{2i})$, where $\mathbf{X}_{1i}$ and $\mathbf{Y}_{1i}$ correspond to missing observations, and $\mathbf{X}_{2i}$ and $\mathbf{Y}_{2i}$ correspond to available observations. (In this final partitioning, the dimensions of the various pieces may vary with i, and elements may be permuted differently for each i according to the pattern of missing values in each observation.) We note that the complete-data sufficient statistics for the parameters in this model are

$$\begin{aligned} N_{\mathbf{x}} &= \Sigma_{i=1}^{n}\mathbf{I}(\mathbf{X}_i = \mathbf{x}), \\ \mathbf{S}_{\mathbf{x}} &= \Sigma_{i=1}^{n}\mathbf{I}(\mathbf{X}_i = \mathbf{x})\mathrm{Y}_i, \text{ and} \qquad (\mathbf{x} \in \Psi) \\ \mathbf{SS}_{\mathbf{x}} &= \Sigma_{i=1}^{n}\mathbf{I}(\mathbf{X}_i = \mathbf{x})\mathbf{Y}_i\mathbf{Y}_i^{T}, \end{aligned} \tag{9}$$

and that the MLEs for the parameters in this model based on the complete data are

$$\hat{p}_{\mathbf{x}} = N_{\mathbf{x}}/n,$$
$$\hat{\mu}_{\mathbf{x}} = \mathbf{S}_{\mathbf{x}}/N_{\mathbf{x}}, \text{ and} \qquad (\mathbf{x} \in \Psi) \tag{10}$$
$$\hat{\Sigma}_{\mathbf{x}} = \mathbf{SS}_{\mathbf{x}}/N_{\mathbf{x}} - \hat{\mu}_{\mathbf{x}}\hat{\mu}_{\mathbf{x}}^{T}.$$

The M-step in this setting simply amounts to evaluating each of the pieces of (10). The main computational burden lies in computing the conditional expectations of the complete data sufficient statistics given the available data under current parameter estimates in each iteration (the E-step).

In the E-step, we wish to compute (under the distribution defined by $\hat{\theta}^{(k)}$)

$$E[N_{\mathbf{x}} \mid \{(\mathbf{X}_{2i}, \mathbf{Y}_{2i}), i = 1, n\}] = \Sigma_{i=1}^{n} E[\mathbf{I}(\mathbf{X}_i = \mathbf{x}) \mid (\mathbf{X}_{2i}, \mathbf{Y}_{2i})],$$
$$E[\mathbf{S}_{\mathbf{x}} \mid \{(\mathbf{X}_{2i}, \mathbf{Y}_{2i}), i = 1, n\}] = \Sigma_{i=1}^{n} E[\mathbf{I}(\mathbf{X}_i = \mathbf{x})\mathbf{Y}_i \mid (\mathbf{X}_{2i}, \mathbf{Y}_{2i})], \text{ and} \quad (\mathbf{x} \in \Psi) \tag{11}$$
$$E[\mathbf{SS}_{\mathbf{x}} \mid \{(\mathbf{X}_{2i}, \mathbf{Y}_{2i}), i = 1, n\}] = \Sigma_{i=1}^{n} E[\mathbf{I}(\mathbf{X}_i = \mathbf{x})\mathbf{Y}_i\mathbf{Y}_i^{T} \mid (\mathbf{X}_{2i}, \mathbf{Y}_{2i})].$$

This computation is facilitated by the following identities:

$$E[\mathbf{I}(\mathbf{X} = \mathbf{x})h(\mathbf{Y}) \mid (\mathbf{X}_2, \mathbf{Y}_2)] = \Pr[\mathbf{X} = \mathbf{x} \mid (\mathbf{X}_2, \mathbf{Y}_2)] \bullet E[h(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}, \mathbf{Y}_2] \tag{12}$$

$$\Pr[\mathbf{X} = \mathbf{x} \mid (\mathbf{X}_2, \mathbf{Y}_2)] = \frac{\mathbf{I}(\mathbf{X}_2 = \mathbf{x}_2)p_{\mathbf{x}}\text{MVN}(\mathbf{y}_2 \mid \mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})}{\sum\limits_{\tilde{\mathbf{x}} \in \Psi} \mathbf{I}(\tilde{\mathbf{x}}_2 = \mathbf{X}_2)p_{\tilde{\mathbf{x}}}\text{MVN}(\mathbf{y}_2 \mid \mu_{\tilde{\mathbf{x}}}, \Sigma_{\tilde{\mathbf{x}}})} \tag{13}$$

$$E[\mathbf{Y}_1 \mid \mathbf{X} = \mathbf{x}, \mathbf{Y}_2] = \mu_{\mathbf{x}}^{(1)} + \Sigma_{\mathbf{x}}^{(12)}\{\Sigma_{\mathbf{x}}^{(22)}\}^{-1}(\mathbf{Y}_2 - \mu_{\mathbf{x}}^{(2)}) \tag{14}$$

$$E[\mathbf{Y}_2 \mid \mathbf{X} = \mathbf{x}, \mathbf{Y}_2] = \mathbf{Y}_2 \tag{15}$$

$$E[\mathbf{Y}_1\mathbf{Y}_i^{T} \mid \mathbf{X} = \mathbf{x}, \mathbf{Y}_2] = \tag{16}$$
$$\Sigma_{\mathbf{x}}^{(11)} - \Sigma_{\mathbf{x}}^{(12)}\{\Sigma_{\mathbf{x}}^{(22)}\}^{-1}\Sigma_{\mathbf{x}}^{(21)} + E[\mathbf{Y}_1 \mid \mathbf{X} = \mathbf{x}, \mathbf{Y}_2]E[\mathbf{Y}_1 \mid \mathbf{X} = \mathbf{x}, \mathbf{Y}_2]^{T}$$

$$E[\mathbf{Y}_1\mathbf{Y}_2^{T} \mid \mathbf{X} = \mathbf{x}, \mathbf{Y}_2] = E[\mathbf{Y}_1 \mid \mathbf{X} = \mathbf{x}, \mathbf{Y}_2] \bullet \mathbf{Y}_2^{T} \tag{17}$$

$$E[\mathbf{Y}_2\mathbf{Y}_2^T \mid \mathbf{X} = \mathbf{x}, \mathbf{Y}_2] = \mathbf{Y}_2\mathbf{Y}_2^T \tag{18}$$

Here, $\mu_{\mathbf{x}}^{(1)}$, $\mu_{\mathbf{x}}^{(2)}$, $\Sigma_{\mathbf{x}}^{(11)}$, $\Sigma_{\mathbf{x}}^{(12)}$, $\Sigma_{\mathbf{x}}^{(21)}$, and $\Sigma_{\mathbf{x}}^{(22)}$ are appropriate partitions of $\mu_{\mathbf{x}}$ and $\Sigma_{\mathbf{x}}$ corresponding to the missing and available parts of **Y**. Equations (14) - (18) are used to estimate the missing parts of $E[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, \mathbf{Y}_2]$ and $E[\mathbf{Y}\mathbf{Y}^T \mid \mathbf{X} = \mathbf{x}, \mathbf{Y}_2]$. Computation of the expectations in (11) is then carried out as follows: For each observation in the data set, compute $E[\mathbf{I}(\mathbf{X}_i = \mathbf{x}) \mid (\mathbf{X}_{2i}, \mathbf{Y}_{2i})]$, $E[\mathbf{I}(\mathbf{X}_i = \mathbf{x})\mathbf{Y}_i \mid (\mathbf{X}_{2i}, \mathbf{Y}_{2i})]$, and $E[\mathbf{I}(\mathbf{X}_i = \mathbf{x})\mathbf{Y}_i\mathbf{Y}_i^T \mid (\mathbf{X}_{2i}, \mathbf{Y}_{2i})]$ for each $\mathbf{x} \in \Psi$ via Equations (12) - (18). Accumulate these over all observations to obtain (11).

In the case of continuous variables only, (MGW) used estimates of parameters based on substituting means for missing observations as initial estimates in the iterative process. This becomes more complicated in the presence of categorical variables. To simplify initialization, we use "blind initialization": we initialize each $p_{\mathbf{x}}$ with 1/r, each $\mu_{\mathbf{x}}$ with **0**, and each $\Sigma_{\mathbf{x}}$ with **I**. Experience so far indicates that the first iteration of the EM algorithm substantially alters the parameter estimates to something comparable to "mean substitution," if that means anything in this context. In any case, this initialization procedure has worked adequately so far in simulation studies.

Having evaluated the MLEs using the EM algorithm, we need a method for evaluating the maximized log-likelihood functions in Equation (6). The likelihood function for the available data is the product of the likelihoods of the available parts of each observation. The likelihood of the available part of a single observation is the marginal density for (X_2, Y_2). This may be obtained from the density for (X, Y) by integrating out X_1 and Y_1. This gives

$$f_{X_2,Y_2}(\mathbf{x}_2,\mathbf{y}_2) = \sum_{\tilde{\mathbf{x}} \in \Psi} \mathbf{I}(\mathbf{x}_2 = \tilde{\mathbf{x}}_2)\, p_{\tilde{\mathbf{x}}}\, \mathrm{MVN}(\mathbf{y}_2 \mid \mu_{\tilde{\mathbf{x}}}, \Sigma_{\tilde{\mathbf{x}}}). \tag{19}$$

The maximized log-likelihood of the available data in a sample is obtained by accumulating the values of the log of (19) over all observations in the sample. Thus, we may evaluate each of the pieces λ_{01}, λ_{02}, λ_{11}, and λ_{12} in Equation (6), from which we may evaluate the test statistic $\lambda = \log(\mathrm{LR})$ given in (5).

As will be seen in the simulation results of Section 6, the FULL Algorithm has one major flaw that must be addressed. That is, it can only be used in cases in which there is adequate representation in each cell to obtain a full-rank estimate of $\Sigma_{\mathbf{x}}$ for each $\mathbf{x} \in \Psi$ in both populations. If r is large, *i.e.*, if there are a large number of categorical variables, or a large number of categories within some categorical variables, or both, then the training samples may need to be extremely large so that all parameters can be estimated accurately. In practice, such large samples may not be available, and it becomes necessary to impose further constraints on the parameters of the model so that the number of parameters required is reduced. This leads us into our discussion of the next algorithm.

5. The COMMON Algorithm

This last algorithm is very similar to the FULL Algorithm except that in our model for the data, we assume that the conditional covariance matrix for the continuous part given the discrete part is common for all $\mathbf{x} \in \Psi$. That is, the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is multivariate normal with mean $\mu_{\mathbf{x}}$ and covariance matrix and Σ not depending on $\mathbf{x}$. Hence, $\theta = \{p_{\mathbf{x}}, \mu_{\mathbf{x}}, \Sigma; \mathbf{x} \in \Psi\}$. This reduces the number of parameters that need to be estimated considerably, and makes parameter estimation possible when some cells are sparse, or not represented at all. We allow the possibility of different parameters for each of the two populations, but within each population, Σ is common across all multinomial cells. This model gives precisely the general location model of Olkin and Tate (1961). The EM algorithm for this model is developed by Little and Schluchter (1985), and they point out that this can be used to implement the GLRT procedure proposed by Krzanowski (1982). What follows is precisely this procedure, with the added feature that we bootstrap the distribution of the test statistic in order to choose critical values to control the P(2|1) error rate. Although Little and Schluchter (1985) describe the EM algorithm for this model in considerable detail, we present a description of the algorithm here that is consistent with the notation of Section 4.

First, we observe that the complete-data sufficient statistics for the parameters in this model are

$$\begin{aligned} N_{\mathbf{x}} &= \Sigma_{i=1}^{n}\mathbf{I}(\mathbf{X}_i = \mathbf{x}), \\ \mathbf{S}_{\mathbf{x}} &= \Sigma_{i=1}^{n}\mathbf{I}(\mathbf{X}_i = \mathbf{x})Y_i, \text{ and} \qquad (\mathbf{x} \in \Psi) \\ \mathbf{SS} &= \Sigma_{i=1}^{n}\mathbf{Y}_i\mathbf{Y}_i^T, \end{aligned} \tag{20}$$

and that the MLEs for the parameters in this model based on the complete data are

$$\begin{aligned} \hat{p}_{\mathbf{x}} &= N_{\mathbf{x}}/n, \\ \hat{\mu}_{\mathbf{x}} &= \mathbf{S}_{\mathbf{x}}/N_{\mathbf{x}}, \text{ and} \qquad (\mathbf{x} \in \Psi), \\ \hat{\Sigma} &= \Sigma_{\mathbf{x}\in\Psi}\hat{p}_{\mathbf{x}}\hat{\Sigma}_{\mathbf{x}}, \end{aligned} \tag{21}$$

where $\hat{\Sigma}_{\mathbf{x}}$ is given by equations (9) and (10). In other words, the MLE of Σ in this case is precisely a weighted average of MLEs of $\Sigma_{\mathbf{x}}$ for each $\mathbf{x}$ based on the FULL model with weights $\hat{p}_{\mathbf{x}}$. Hence, we may perform the M-step in this algorithm with exactly the same formulas as the M-step in the FULL Algorithm, except that after each $\hat{\Sigma}_{\mathbf{x}}$ is computed, we average these according to (21) to obtain the updated estimate of the common Σ. The E-step for this algorithm is also identical to the E-step in the FULL Algorithm, except that throughout formulas (12) - (18), each $\hat{\Sigma}_{\mathbf{x}}$ is replaced by the common $\hat{\Sigma}$. The evaluation of the maximized log-likelihood functions in (6) is also performed using (19), as in the FULL algorithm, with again, the only difference being that each $\hat{\Sigma}_{\mathbf{x}}$ is replaced by the common $\hat{\Sigma}$.

6. Simulation Results

We have performed simulations of each of the three algorithms (INDICATOR, FULL, and COMMON) based on several different parameter configurations in order to determine how well the algorithm controls the Type I misclassification probability as desired, and to assess the power P(2|2) of each algorithm. We also keep track of how many times the algorithm fails to classify the observation at all. These failures occur when for some reason the simulated data fails to yield full-rank estimates of all required covariance matrix parameters. This results in an undefined test statistic λ. This happens most frequently in the FULL algorithm, and is caused by a very few number of observations falling into one or more of the multinomial cells. It happens occasionally in the INDICATOR algorithm when at least one possible value of a categorical variable is not represented. Failures may occur when the test statistic is undefined for the sample which we are trying to classify, and also when the statistic is undefined for attempted bootstrap samples. We see in our simulations that the COMMON algorithm is least susceptible to these types of failures.

Our first simulation involved the same parameter configurations used in BGWMF (Case 3: Mixture of Categorical and Continuous Variables). That is, we consider the case in which the categorical part is a single Bernoulli variable and the continuous part is a single normal random variable independent of the categorical variable. For population π_1, the Bernoulli parameter is $p_1 = 0.1$. The mean and variance of the continuous variable are $\mu_1 = 0$ and $\sigma_1^2 = 0.5$, respectively. For population π_2, we use $p_2 = 0.9$, 0.7, and 0.5, $\sigma_2^2 = 1.0$, and $\mu_2 = 0.5 + \Delta\sigma_2^2$ where Δ takes on values 0, 1, 2, and 3. The observed significance level $\hat{P}(2|1)$ is the proportion of times out of 500 simulated trials in which the variable **V** is classified into π_2 when, in fact, it was simulated from π_1. The estimated power $\hat{P}(2|2)$ is the proportion of times out of 500 simulated trials in which the variable **V** is classified into π_2 when, in fact, it was simulated from π_2. In order to achieve an approximate significance level of $\alpha = 0.05$, the variable **V** was classified into

π_2 if the test statistic λ is less than or equal to λ_α^*, the 0.05-th empirical quantile of $\{\lambda_1^*, \lambda_2^*, \ldots, \lambda_B^*\}$.

For $N_1 = N_2 = 50$ and $B = 99$, the power estimates are plotted in Figure 5, based on simulations with no missing data. We see that the FULL and COMMON algorithms agree very well with the power curves plotted in BGWMF (Figure 2). In fact, with no missing data, the COMMON algorithm is essentially equivalent to the method of BGWMF, so these simulation results should agree very well, as they do. The INDICATOR algorithm does not agree well with the FULL and COMMON algorithms. For this reason, the points corresponding to the INDICATOR algorithm are not connected with lines, since this would clutter the plot. It would seem that the INDICATOR algorithm has higher power in general than the other two. This is surprising since this algorithm does not model well the true distribution of the binary variable. A closer examination of the simulation results shows that this is, in fact, misleading, since the INDICATOR tends to yield a significance level nearly twice the desired 0.05 level. This can be seen in Figure 6, which shows the power estimate plotted versus the observed significance level. Each plot in Figure 6 corresponds to a specific value of Δ. We can also see in Figure 6 that the COMMON algorithm most accurately achieves the desired $\alpha = 0.05$ significance level.

In Figures 7 and 8, we show corresponding plots based on data with missing values. In these simulations, each variable in each observation was deleted independently with probability 0.1, so that roughly 10% of the data is missing. We see an overall decrease in the power of all three algorithms compared to the full-data case, but this is to be expected since the test is based on less available data. Otherwise, the results of the missing-data case are comparable to the results of the full-data case.

We have tabulated the results of this simulation in Table 1. The ERROR column shows the percentage of times out of the 500 simulations that the algorithm failed to classify **V** due to singular parameter estimates. We see that the FULL algorithm is most

NO MISSING DATA

PROG	DELTA	P2	SIG	PWR	ERR
I	0	0.9	0.1006	0.8893	0.6
I	0	0.7	0.0946	0.6881	0.6
I	0	0.5	0.0946	0.5191	0.6
I	1	0.9	0.0926	0.8994	0.6
I	1	0.7	0.0905	0.8089	0.6
I	1	0.5	0.0885	0.7183	0.6
I	2	0.9	0.0885	0.9638	0.6
I	2	0.7	0.0885	0.9376	0.6
I	2	0.5	0.0885	0.9175	0.6
I	3	0.9	0.0845	0.9980	0.6
I	3	0.7	0.0825	0.9940	0.6
I	3	0.5	0.0825	0.9920	0.6
F	0	0.9	0.0681	0.6085	6.0
F	0	0.7	0.0638	0.4938	2.8
F	0	0.5	0.0779	0.3668	2.4
F	1	0.9	0.0581	0.7828	7.0
F	1	0.7	0.0723	0.7190	3.2
F	1	0.5	0.0717	0.6721	2.4
F	2	0.9	0.0746	0.9616	6.2
F	2	0.7	0.0825	0.9340	3.0
F	2	0.5	0.0799	0.9139	2.4
F	3	0.9	0.0423	0.9937	5.4
F	3	0.7	0.0887	0.9938	3.0
F	3	0.5	0.0799	0.9877	2.4
C	0	0.9	0.0480	0.5920	0.0
C	0	0.7	0.0480	0.4640	0.0
C	0	0.5	0.0460	0.3660	0.0
C	1	0.9	0.0520	0.8540	0.0
C	1	0.7	0.0500	0.7500	0.0
C	1	0.5	0.0480	0.6920	0.0
C	2	0.9	0.0460	0.9640	0.0
C	2	0.7	0.0480	0.9440	0.0
C	2	0.5	0.0380	0.9240	0.0
C	3	0.9	0.0500	1.0000	0.0
C	3	0.7	0.0500	0.9940	0.0
C	3	0.5	0.0400	0.9940	0.0

10 PERCENT MISSING DATA

PROG	DELTA	P2	SIG	PWR	ERR
I	0	0.9	0.0872	0.8661	1.4
I	0	0.7	0.0884	0.6406	0.4
I	0	0.5	0.0823	0.4940	0.4
I	1	0.9	0.0791	0.8905	1.4
I	1	0.7	0.0843	0.7952	0.4
I	1	0.5	0.0763	0.6888	0.4
I	2	0.9	0.0791	0.9513	1.4
I	2	0.7	0.0805	0.9155	0.6
I	2	0.5	0.0803	0.8755	0.4
I	3	0.9	0.0730	0.9858	1.4
I	3	0.7	0.0763	0.9799	0.4
I	3	0.5	0.0763	0.9779	0.4
F	0	0.9	0.0561	0.5611	0.2
F	0	0.7	0.0481	0.4609	0.2
F	0	0.5	0.0481	0.3507	0.2
F	1	0.9	0.0601	0.7896	0.2
F	1	0.7	0.0501	0.6954	0.2
F	1	0.5	0.0561	0.6353	0.2
F	2	0.9	0.0520	0.9420	0.0
F	2	0.7	0.0500	0.8920	0.0
F	2	0.5	0.0540	0.8660	0.0
F	3	0.9	0.0582	0.9799	0.4
F	3	0.7	0.0501	0.9780	0.2
F	3	0.5	0.0541	0.9739	0.2
C	0	0.9	0.0440	0.5700	0.0
C	0	0.7	0.0480	0.4520	0.0
C	0	0.5	0.0520	0.3520	0.0
C	1	0.9	0.0360	0.8080	0.0
C	1	0.7	0.0440	0.7140	0.0
C	1	0.5	0.0440	0.6500	0.0
C	2	0.9	0.0440	0.9420	0.0
C	2	0.7	0.0440	0.8980	0.0
C	2	0.5	0.0440	0.8640	0.0
C	3	0.9	0.0440	0.9840	0.0
C	3	0.7	0.0460	0.9760	0.0
C	3	0.5	0.0440	0.9780	0.0

PROGRAM CODE: I - INDICATOR F - FULL C - COMMON

Table 1. Listing of results from first simulation.

susceptible to failures of this sort, failing as much as 6 to 7% of the time when $p_2 = 0.9$ and no data is missing. The INDICATOR algorithm failed somewhat less frequently, and the COMMON algorithm never failed in this study.

In our next simulation study, we consider the case in which we have two categorical variables, each with two categories, and two continuous variables. For population π_1, each possible combination of the categorical part ($\mathbf{X}$ = (1,1), (1,2), (2,1) and (2,2)) occurs with probability 1/4. The conditional distribution of the continuous part is MVN($\mathbf{0}$, Σ_1), where

$$\Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \tag{22}$$

within each multinomial cell (*i.e.*, conditional on each possible value of the discrete part). For population π_2, the conditional covariance matrix for the continuous part is Σ_2, where Σ_2 is given by

$$\Sigma_2 = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}, \tag{23}$$

We use three different probability distributions for the discrete part, and four different configurations of mean vectors for the conditional distributions of the continuous part given each possible discrete part. In the plots and tables which follow, the three probability distributions are coded with the variable PCODE, which takes on values 1, 2, and 3. The four mean vector configurations are coded with the variable MCODE, which takes on values 1, 2, 3, and 4. The parameter configurations defined by these codes are shown in Table 2.

PCODE	$\Pr[\mathbf{X} = (1, 1)]$	$\Pr[\mathbf{X} = (1, 2)]$	$\Pr[\mathbf{X} = (2, 1)]$	$\Pr[\mathbf{X} = (2, 2)]$
1	0.25	0.25	0.25	0.25
2	0.50	0.20	0.20	0.10
3	0.80	0.10	0.10	0.00
MCODE	$E[\mathbf{Y}\vert\mathbf{X} = (1, 1)]$	$E[\mathbf{Y}\vert\mathbf{X} = (1, 2)]$	$E[\mathbf{Y}\vert\mathbf{X} = (2, 1)]$	$E[\mathbf{Y}\vert\mathbf{X} = (2, 2)]$
1	(0,0)	(0,0)	(0,0)	(0,0)
2	(2,2)	(1,1)	(1,1)	(0,0)
3	(0,0)	(1,1)	(1,1)	(2,2)
4	(2,2)	(2,2)	(2,2)	(2,2)

Table 2. Definitions for parameter codes used in out second simulation study.

PCODE = 1 corresponds to a uniform distribution across all multinomial cells. PCODE = 2 and PCODE = 3 correspond to distributions increasingly favoring cell (1,1). MCODE = 1 corresponds to a mean configuration identical to that for population π_1. MCODE = 2 and MCODE = 3 correspond to changes in mean for certain cells, and MCODE = 4 corresponds to the sum of these two changes. For PCODE = 1 and MCODE = 1, population π_2 is identical to population π_1 except for the correlation between the two continuous variables.

As in our first study, we take $N_1 = N_2 = 50$, $B = 99$, $\alpha = 0.05$, and base our observed significance level and power estimates on 500 replications of the procedure in each case. Figure 9 shows the power estimates plotted versus the mean configuration when no data is missing. Figure 10 shows plots of the power estimate versus the observed significance level for each mean configuration. Figures 11 and 12 are corresponding plots for approximately 10% missing data, with data deleted at random in the same manner as our previous study. Table 3 shows a listing of these results, including the percentages of failures due to singular parameter estimates.

In Figure 9, we see the power increases in general as the separation between in means increases (*i.e.*, as MCODE changes from 1 to 4). MCODE = 2 and MCODE = 3 actually correspond to the same degree of difference in means, so the power for these are

NO MISSING DATA

PROG	PCODE	MCODE	SIG	PWR	ERR
I	1	1	0.0400	0.2300	0.0
I	2	1	0.0580	0.2520	0.0
I	3	1	0.1091	0.5859	1.0
I	1	2	0.0640	0.4540	0.0
I	2	2	0.0560	0.5920	0.0
I	3	2	0.0889	0.8283	1.0
I	1	3	0.0460	0.4380	0.0
I	2	3	0.0500	0.3080	0.0
I	3	3	0.1172	0.5657	1.0
I	1	4	0.0720	0.8440	0.0
I	2	4	0.0700	0.8540	0.0
I	3	4	0.0788	0.9010	1.0
F	1	1	0.0480	0.1660	0.0
F	2	1	0.0460	0.1926	8.6
F	3	1	0.0507	0.2230	40.8
F	1	2	0.0541	0.3627	0.2
F	2	2	0.0576	0.5543	9.8
F	3	2	0.0766	0.8029	45.2
F	1	3	0.0441	0.4309	0.2
F	2	3	0.0500	0.3420	9.2
F	3	3	0.0547	0.3029	45.2
F	1	4	0.0700	0.7880	0.0
F	2	4	0.0625	0.8060	7.2
F	3	4	0.0871	0.9024	42.6
C	1	1	0.0480	0.2400	0.0
C	2	1	0.0580	0.2660	0.0
C	3	1	0.0440	0.3360	0.0
C	1	2	0.0380	0.4680	0.0
C	2	2	0.0440	0.6080	0.0
C	3	2	0.0540	0.8180	0.0
C	1	3	0.0580	0.4840	0.0
C	2	3	0.0460	0.3380	0.0
C	3	3	0.0480	0.3400	0.0
C	1	4	0.0580	0.8380	0.0
C	2	4	0.0520	0.8720	0.0
C	3	4	0.0460	0.9180	0.0

10 PERCENT MISSING DATA

PROG	PCODE	MCODE	SIG	PWR	ERR
I	1	1	0.0340	0.1860	0.0
I	2	1	0.0380	0.2220	0.0
I	3	1	0.1084	0.5072	2.2
I	1	2	0.0500	0.3820	0.0
I	2	2	0.0520	0.5360	0.0
I	3	2	0.0757	0.8200	2.2
I	1	3	0.0340	0.3720	0.0
I	2	3	0.0480	0.2540	0.0
I	3	3	0.1207	0.4806	2.2
I	1	4	0.0660	0.7840	0.0
I	2	4	0.0740	0.7980	0.0
I	3	4	0.0838	0.8875	2.2
F	1	1	0.0460	0.1720	0.0
F	2	1	0.0425	0.1680	1.2
F	3	1	0.0602	0.2229	66.8
F	1	2	0.0600	0.3360	0.0
F	2	2	0.0468	0.4725	1.8
F	3	2	0.0637	0.6115	68.6
F	1	3	0.0460	0.3040	0.0
F	2	3	0.0489	0.2688	1.8
F	3	3	0.0600	0.1667	70.0
F	1	4	0.0560	0.7520	0.0
F	2	4	0.0848	0.7475	1.0
F	3	4	0.0325	0.6558	69.2
C	1	1	0.0400	0.1860	0.0
C	2	1	0.0340	0.2100	0.0
C	3	1	0.0400	0.2800	0.0
C	1	2	0.0360	0.4020	0.0
C	2	2	0.0380	0.5620	0.0
C	3	2	0.0560	0.7660	0.0
C	1	3	0.0500	0.4140	0.0
C	2	3	0.0380	0.2840	0.0
C	3	3	0.0400	0.2880	0.0
C	1	4	0.0560	0.7980	0.0
C	2	4	0.0620	0.8100	0.0
C	3	4	0.0560	0.8740	0.0

PROGRAM CODE: I – INDICATOR F – FULL C – COMMON

Table 3. Listing of results from second simulation study.

not expected to be too different. In fact, the power estimates for MCODE = 2 and MCODE = 3 are very similar when PCODE = 1. However, they are not very similar when PCODE = 2 or 3. In this case, power is lower for MCODE = 3 than for MCODE = 2. This results since for MCODE = 3, the means differ in sparse cells, whereas for MCODE = 2, the means differ most in the most common cell (corresponding to **X** = (1,1)), making it more easy to differentiate between the two populations. We see similar patterns in Figure 11, and can also see a general decrease in power due to missing values.

These plots seem to indicate that the INDICATOR and COMMON algorithms have very similar power, these being generally better than the FULL algorithm. As in our first study, we notice in Figures 10 and 12 that the INDICATOR algorithm has a tendency to yield a higher significance level than desired, especially when PCODE = 3 (*i.e.*, when some cells are very sparse).

We see from Table 3 that the INDICATOR algorithm fails occasionally due to singular covariance matrix, especially when PCODE = 3. The FULL algorithm does much worse when cells are sparse. The FULL algorithm fails about two-thirds of the time when PCODE = 3 and data is missing! When some cells occur with very low probability, it is necessary to have very large samples so that each cell is represented enough to obtain a full-rank estimate of the covariance matrix within that cell. Samples of size 50, cells are not adequately represented about 2/3 of the time. Once again, we see that the COMMON algorithm is least susceptible to failures due to singular parameter estimates.

Readers may wonder why the algorithm doesn't fail *every time* for PCODE = 3 since cell (1,1) is *never* represented. If a cell probability is estimated to be zero, the covariance matrix estimate for that cell is never used in the computation of λ, and can therefore be disregarded. It is not cell (1,1) that is the problem here, rather it is cells (0,1) and (1,0). Readers may also find it strange that for PCODE = 2, there are fewer failures in the FULL algorithm when data is missing than when all data is available. This may be explained intuitively as follows: When data is missing in the discrete part, there is some possibility that the observation falls into any of a number of cells. This observation contributes to the parameter estimates for all cells to which the observation might truly belong, resulting in fewer rank problems in sparse cells.

7. Concluding Remarks

In this report, we have extended the results of BGWMF and MGW to perform discriminant analysis with categorical and continuous variables when data is missing. We presented three algorithms for doing so. In simulation studies, we have observed that the INDICATOR algorithm has a tendency to yield a higher Type I error rate than desired. The FULL algorithm often fails due to singular parameter estimates when some value of the discrete part is sparsely represented. The COMMON algorithm seems to avoid these problems, and is therefore the preferred algorithm, especially when samples are small and the assumption of a common covariance matrix across all multinomial cells is reasonable. The code has now been transferred to MRC and Dr. Mark Fisk is applying these techniques to some existing seismic data.

8. References

Baek, J., Gray, H. L., Woodward, W. A., Miller. J. W., and Fisk, M. D. (1994), "A Bootstrap Generalized Likelihood Ratio Test in Discriminant Analysis," submittted to *Computational Statistics and Data Analysis*.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation from Incomplete Data via the *EM* Algorithm" with discussion, *J. Royal Stat. Soc.*, ,1-38.

Efron, B., (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1-26.

Little, R. J. A. & Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R. J. A. & Schluchter, M. D. (1985) Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* 72, 497-512.

Krzanowski, W. J. (1980) Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics* 36, 493-499.

Krzanowski, W. J. (1982) Mixtures of continuous and categorical variables in discriminant analysis: a hypothesis-testing approach. *Biometrics* 38, 991-1002.

Miller, J. W., Gray, H. L., and Woodward, W. A. (1993), "A Bootstrap Generalized Likelihood Ratio Test in Discriminant Analysis."

Olkin, I., & Tate, R. F. (1961) Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Statist.* 32, 448-65

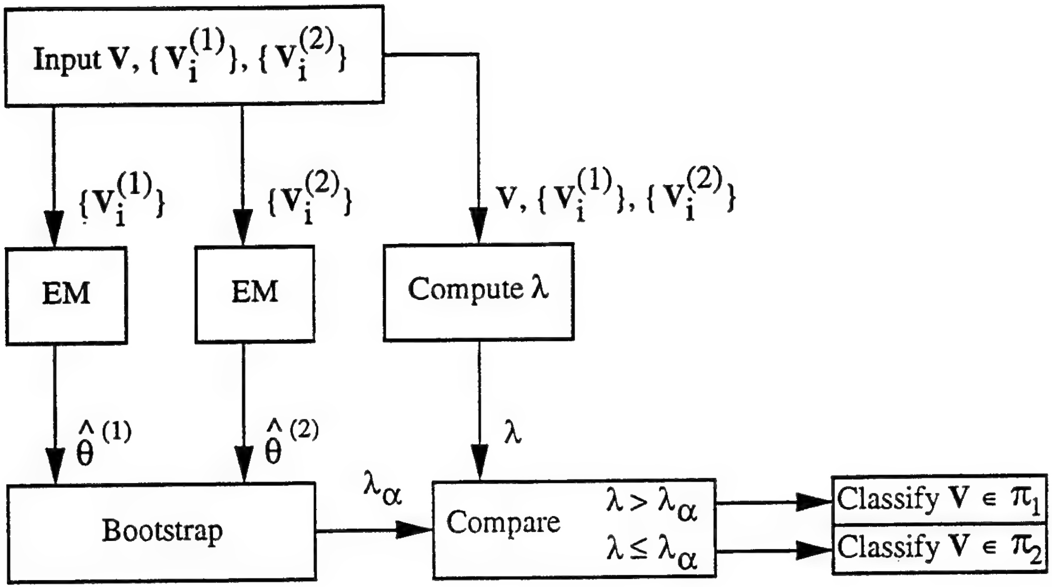


Figure 1. Flowchart for the GLRT procedure for discriminant analysis

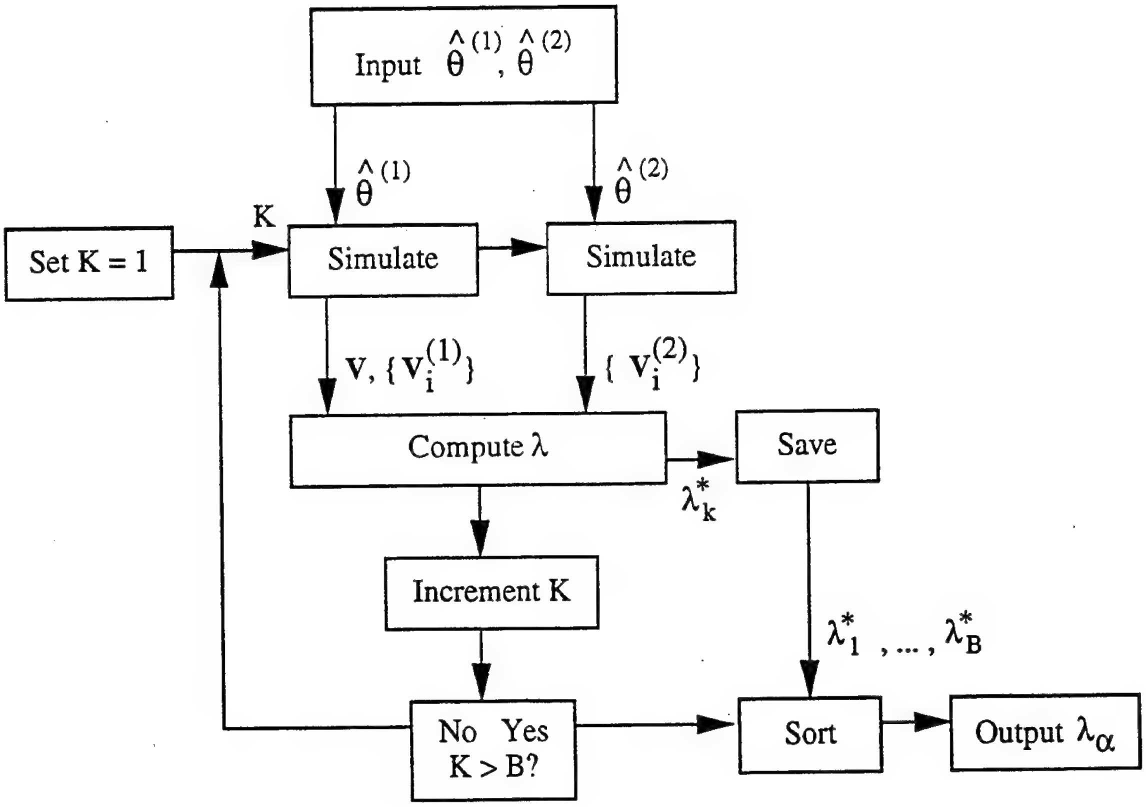


Figure 2. Flowchart for the bootstrapping portion of the GLRT procedure.

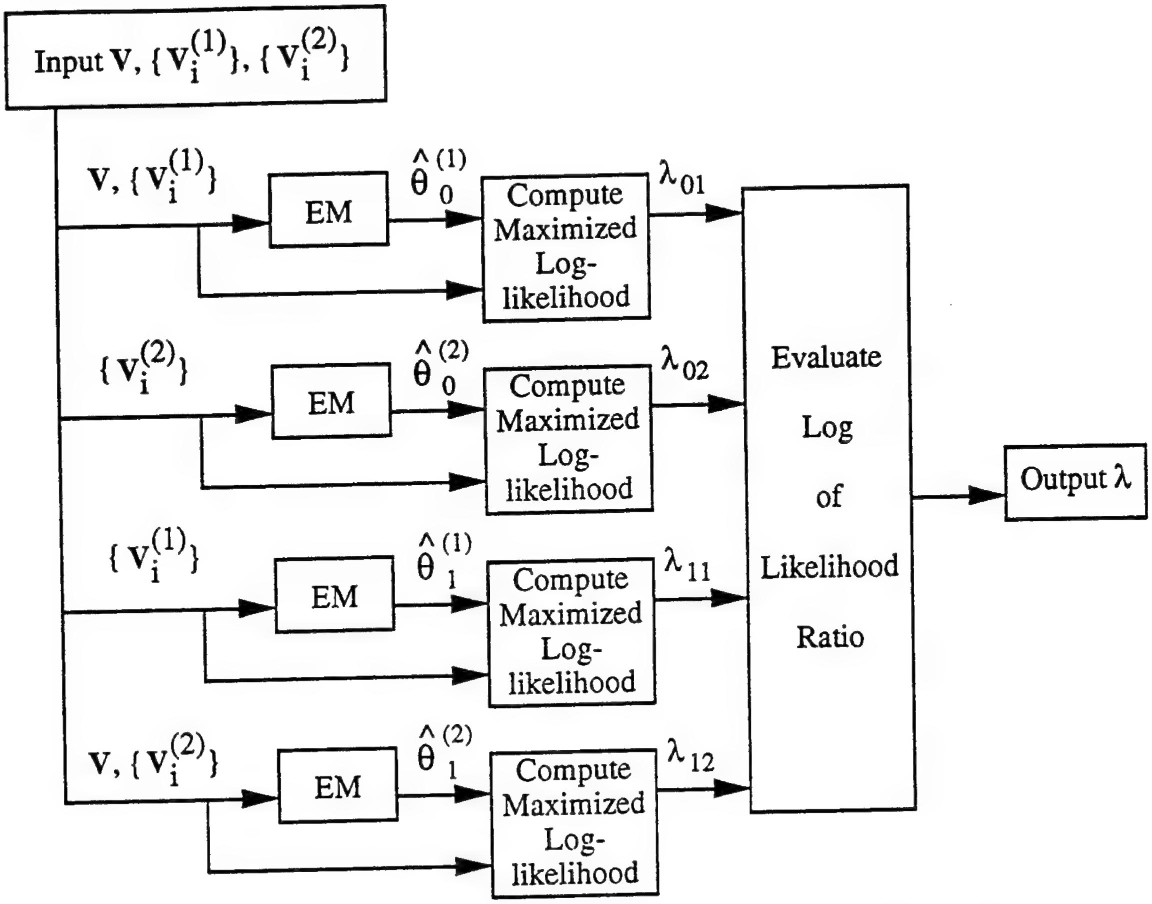


Figure 3. Flowchart for computing the test statistic λ in the GLRT procedure.

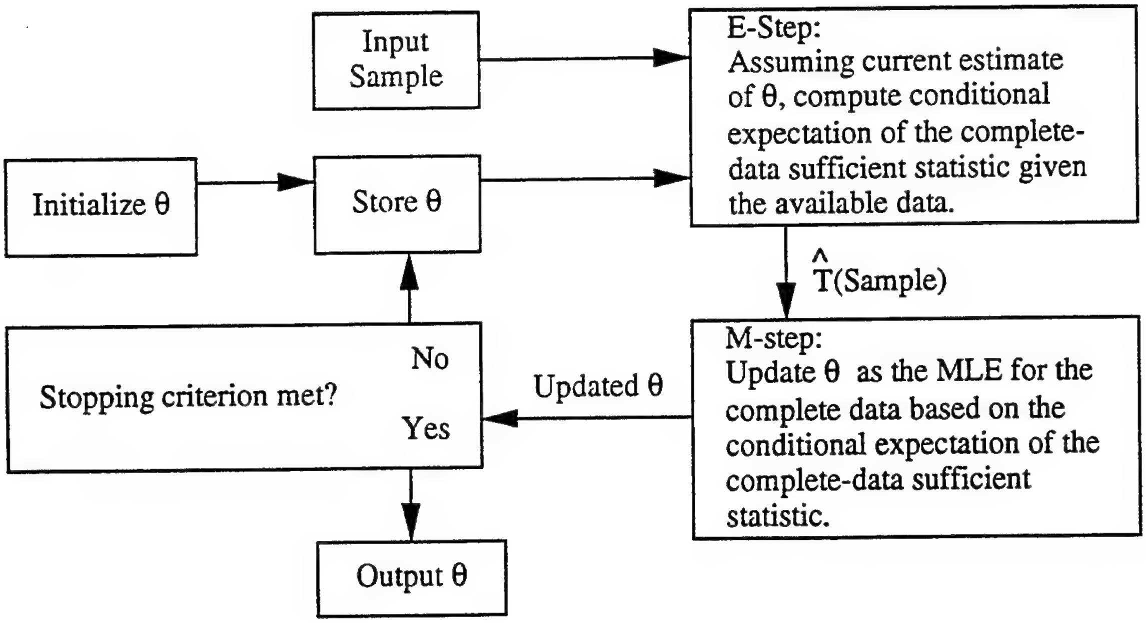

Input Sample
E-Step:
Assuming current estimate of θ, compute conditional expectation of the complete-data sufficient statistic given the available data.
Initialize θ
Store θ
$\hat{T}$(Sample)
Stopping criterion met?
No
Yes
Updated θ
M-step:
Update θ as the MLE for the complete data based on the conditional expectation of the complete-data sufficient statistic.
Output θ

Figure 4. Flowchart for the EM algorithm.

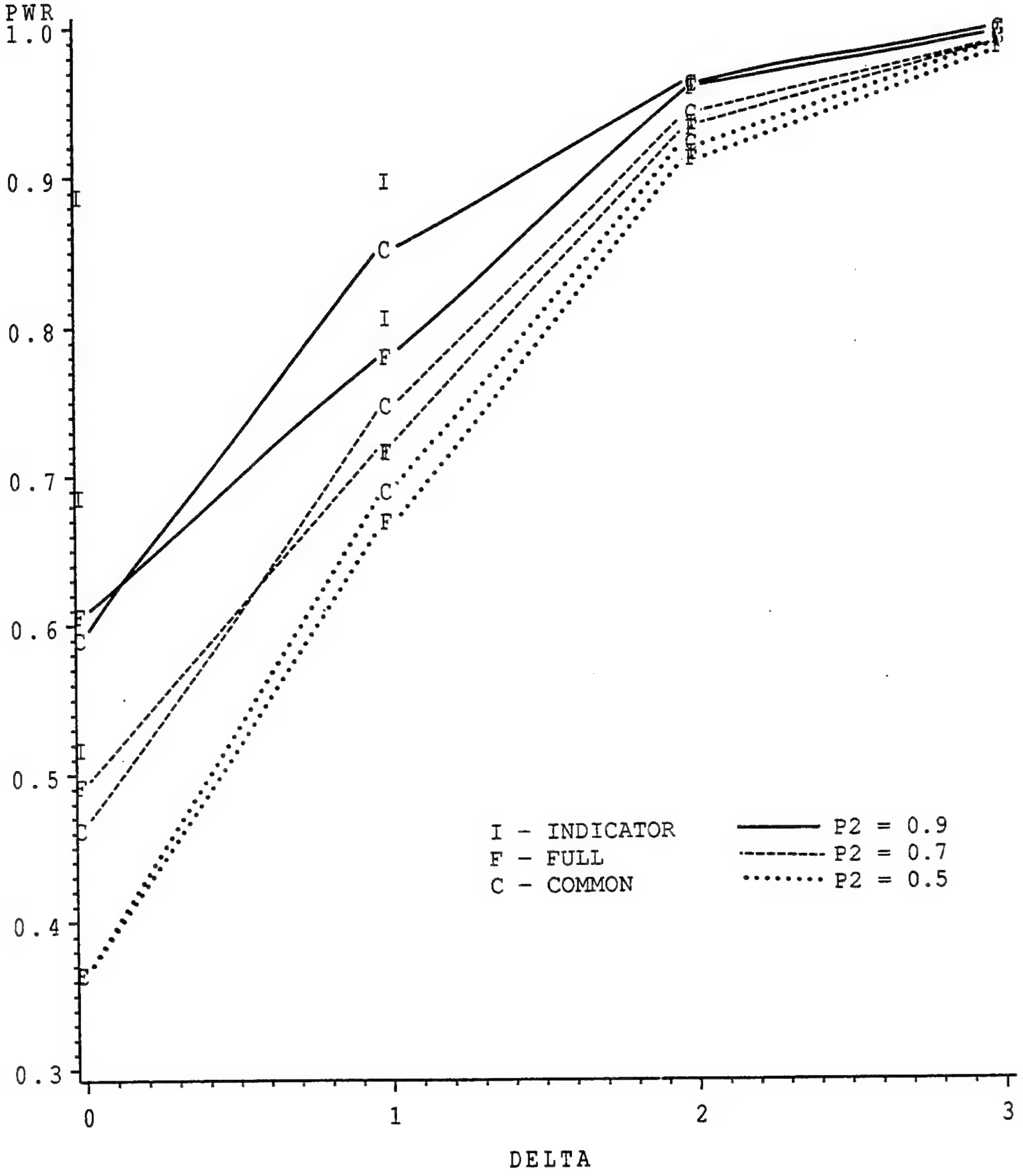
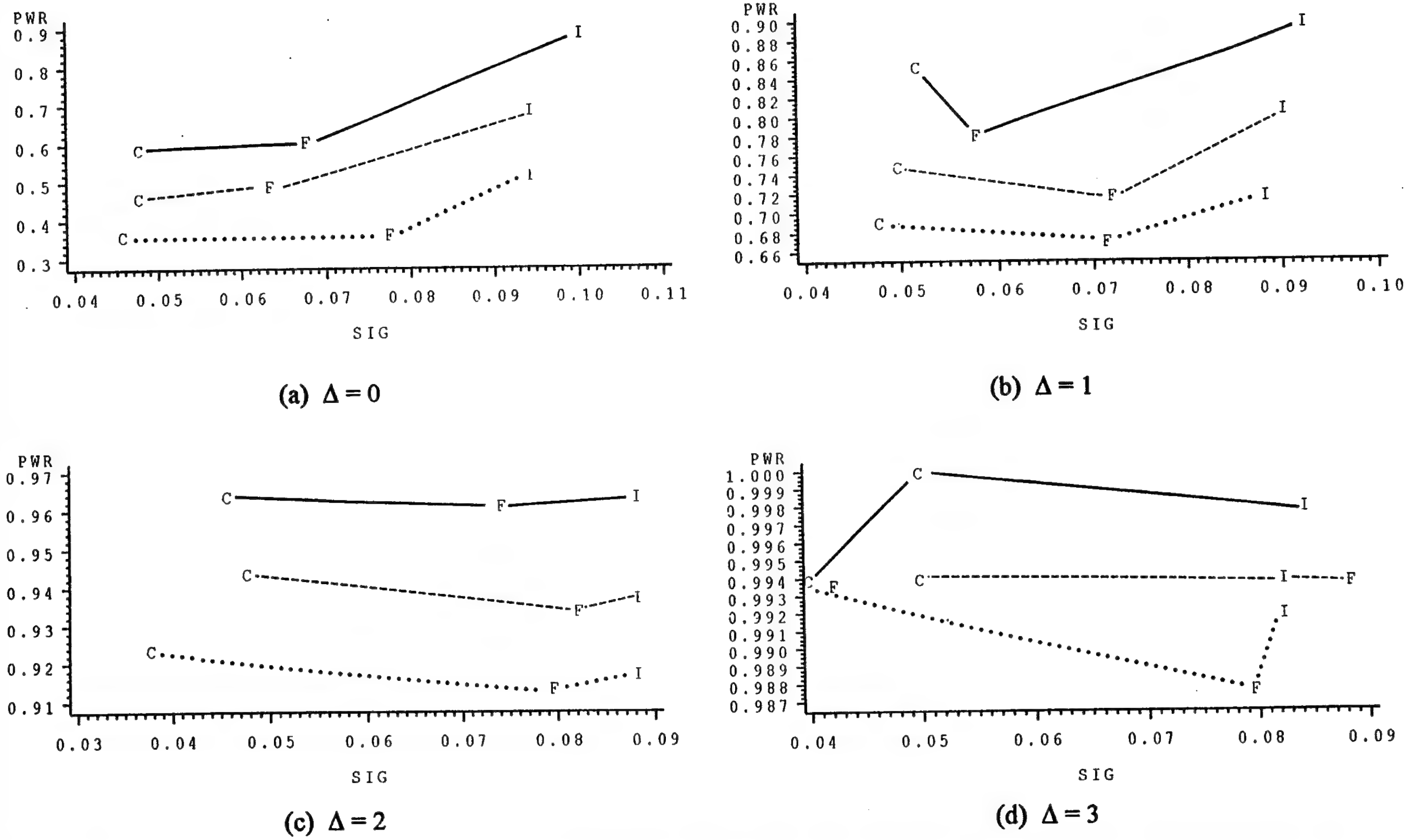


Figure 5. Power estimates when no data is missing for each of the three algorithms, with data modeled as a Bernoulli random variable and an independent normal random variable. Parameters for population π_1 are $p_1 = 0.1$, $\mu_1 = 0$, and $\sigma_1^2 = 0.5$. Power estimates are based on the following configurations for population π_2: $p_2 = 0.9$, 0.7, and 0.5, $\sigma_2^2 = 1.0$, and $\mu_2 = 0.5 + \Delta\sigma_2^2$, where Δ takes on values 0, 1, 2, and 3. For each value of Δ, the symbols I, F, and L are plotted at the corresponding power.



(a) $\Delta = 0$

(b) $\Delta = 1$

(c) $\Delta = 2$

(d) $\Delta = 3$

Figure 6. Plots of estimated power versus observed significance for each value of Δ when no data is missing. Symbol used is I for INDICATOR algorithm, F for FULL, and C for COMMON. The solid line corresponds to $p_2 = 0.9$, the dashed line to $p_2 = 0.7$, and the dotted line to $p_2 = 0.5$. The symbols I, F, and C are plotted at positions associated with the ordered pairs (significance level, power). The lines are not indicative of significance level or power but are only drawn to identify the ordered pairs associated with a particular value of p_2.

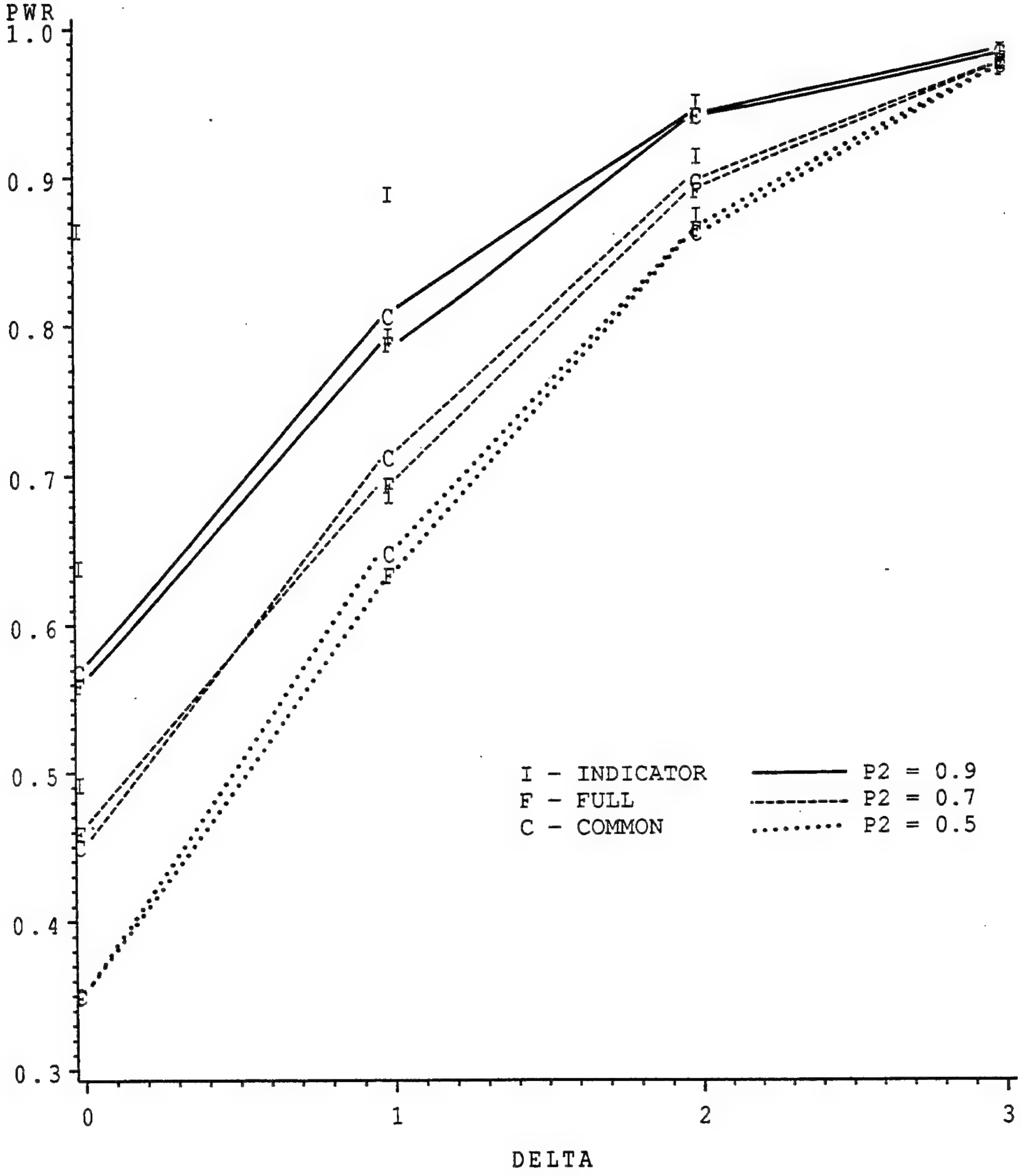
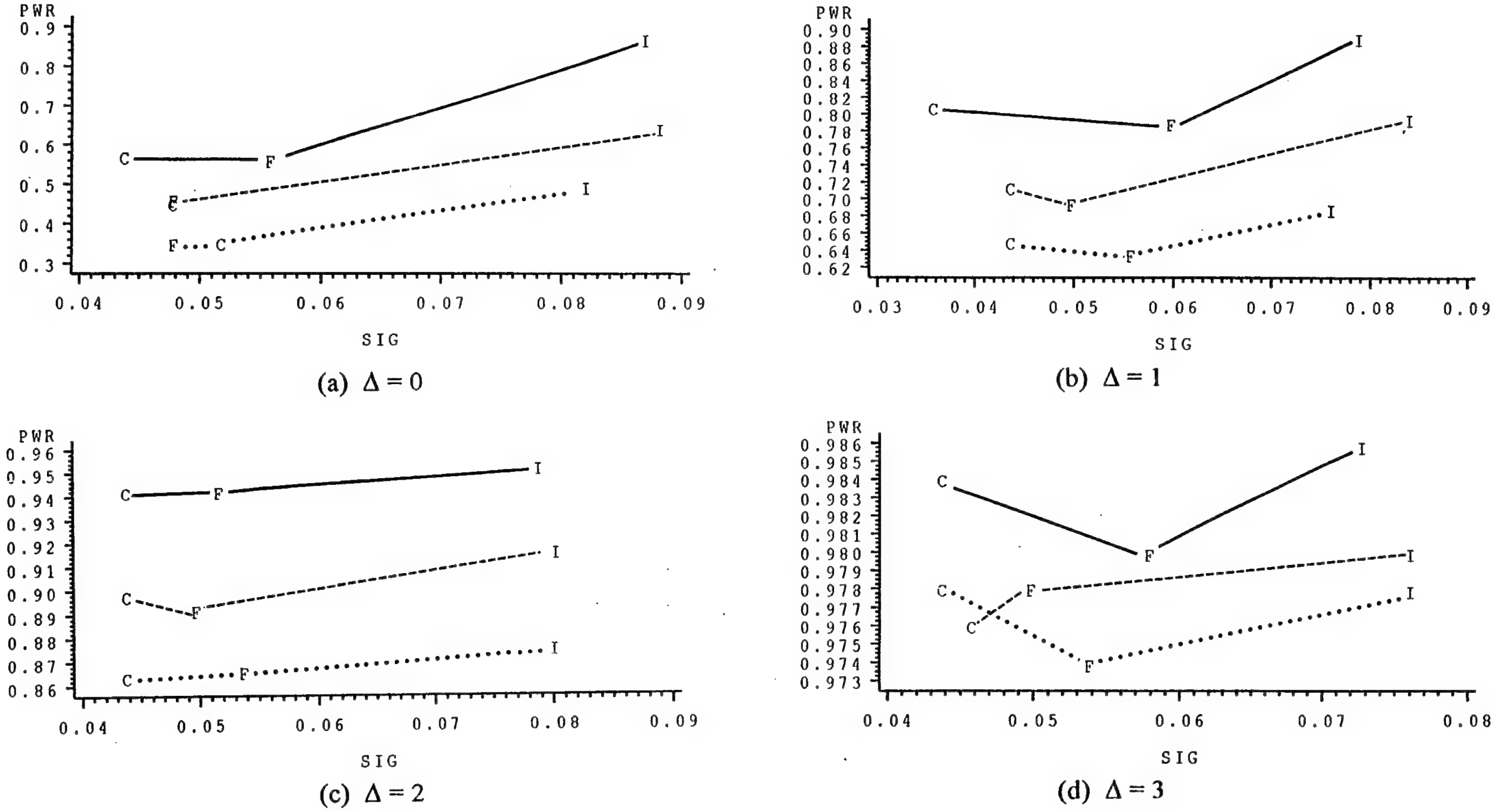


Figure 7. Power estimates when approximately 10% of the data is missing for each of the three algorithms, with data modeled as a Bernoulli random variable and an independent normal random variable. Parameters for population π_1 are $p_1 = 0.1$, $\mu_1 = 0$, and $\sigma_1^2 = 0.5$. Power estimates are based on the following configurations for population π_2: $p_2 = 0.9$, 0.7, and 0.5, $\sigma_2^2 = 1.0$, and $\mu_2 = 0.5 + \Delta\sigma_2^2$, where Δ takes on values 0, 1, 2, and 3.



(a) $\Delta = 0$

(b) $\Delta = 1$

(c) $\Delta = 2$

(d) $\Delta = 3$

Figure 8. Plots of estimated power versus observed significance for each value of Δ when 10% of the data is missing. Symbol used is I for INDICATOR algorithm, F for FULL, and C for COMMON. The solid line corresponds to $p_2 = 0.9$, the dashed line to $p_2 = 0.7$, and the dotted line to $p_2 = 0.5$.

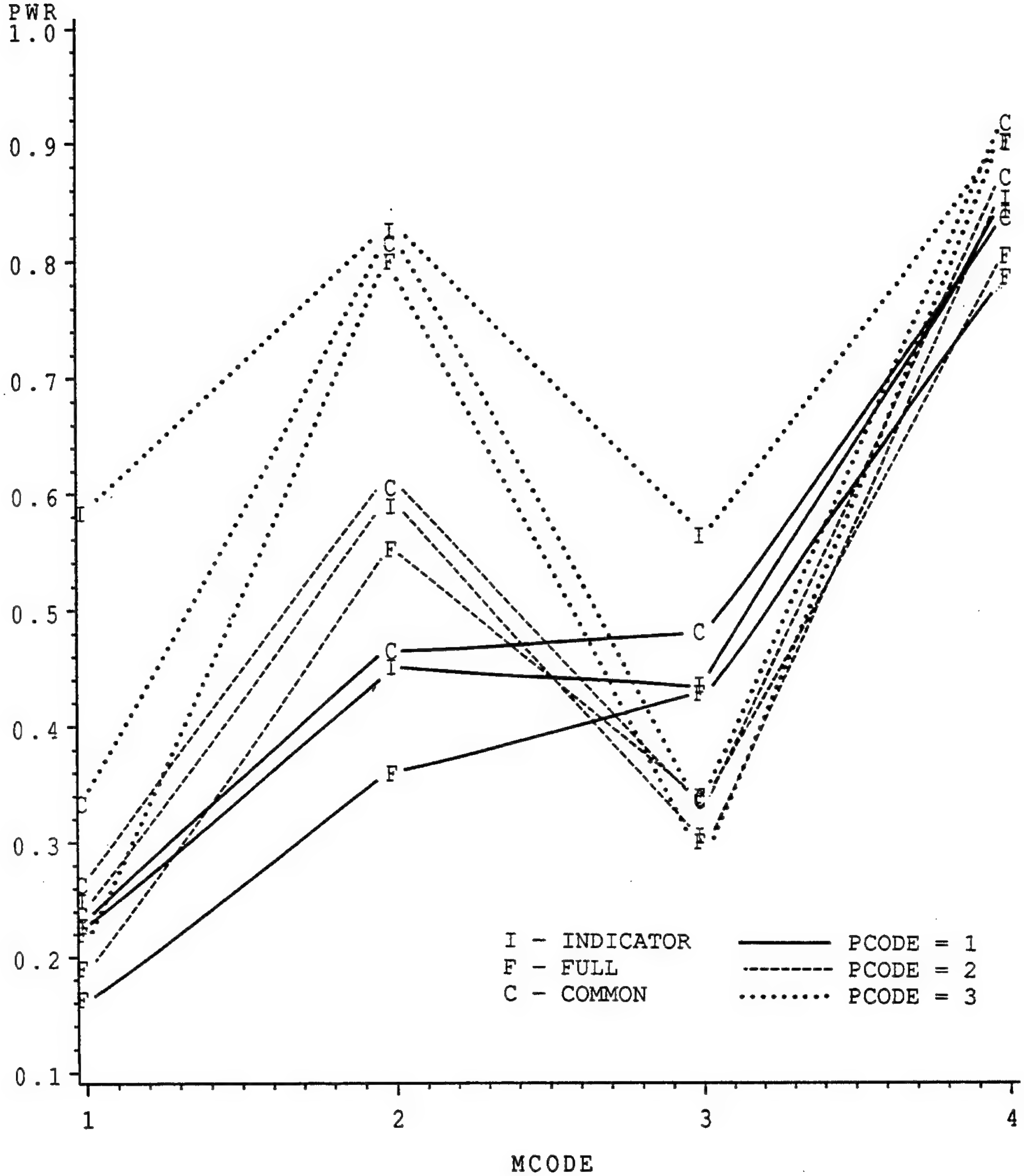
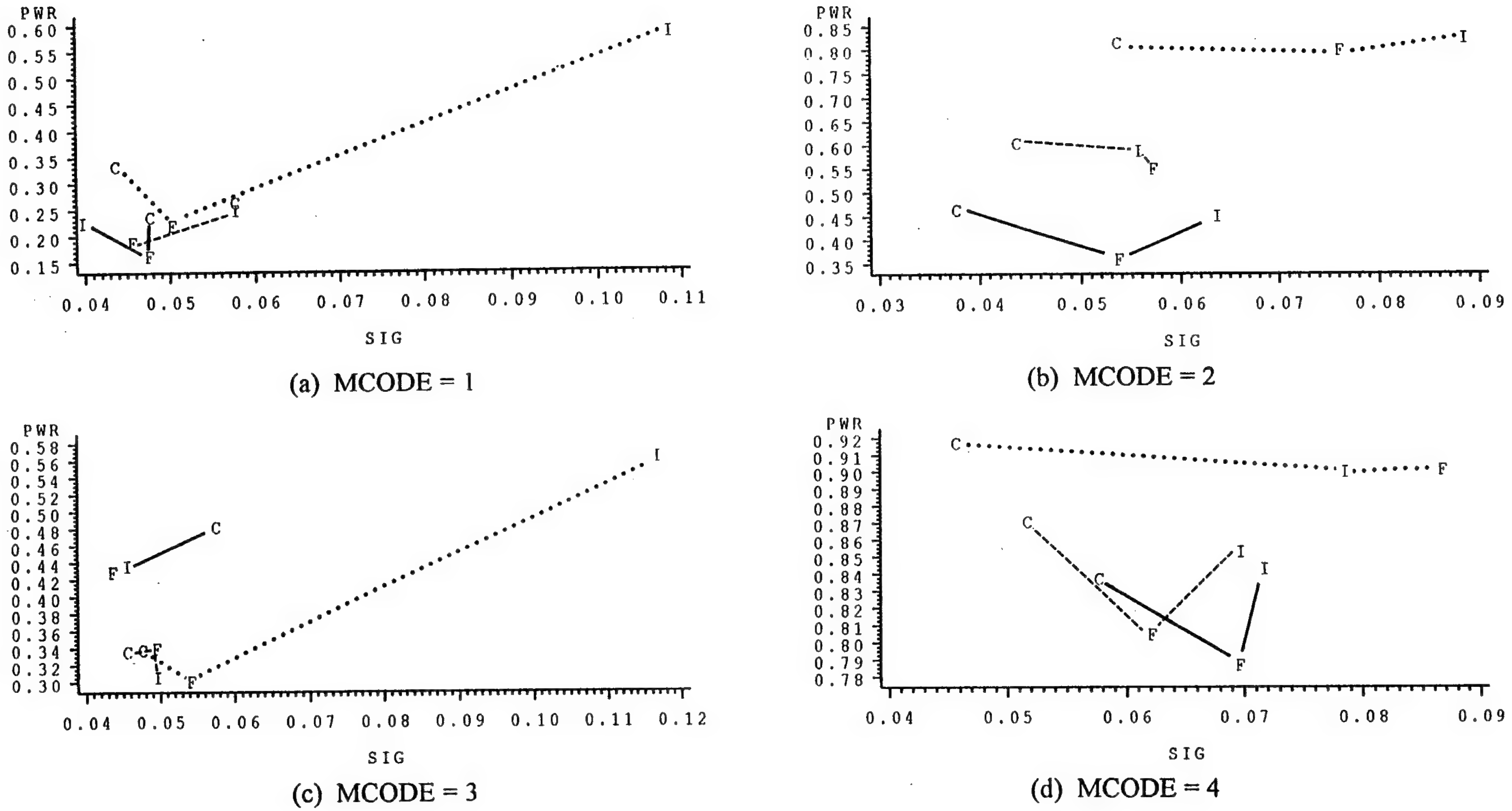


Figure 9. Power estimates when no data is missing for each of the three algorithms when data have two binary and two continuous variates, possibly dependent. For population π_1, each possible combination of the binary part occurs with probability 1/4. The conditional distribution of the continuous part is MVN($\mathbf{0}$, Σ_1), where Σ_1 is a 2x2 matrix with diagonal elements of one and off-diagonal elements of 0.5, within each multinomial cell. For population π_2, the conditional covariance matrix for the continuous part is Σ_2, where Σ_2 has ones on the diagonal and off-diagonal elements of -0.5. Several distributions for the discrete part, and several choices of mean vectors are used, as defined in Table 2.



(a) MCODE = 1

(b) MCODE = 2

(c) MCODE = 3

(d) MCODE = 4

Figure 10. Plots of estimated power versus observed significance for each mean configuration in our second study when no data is missing. Symbol used is I for INDICATOR algorithm, F for FULL, and C for COMMON. The solid line corresponds to PCODE = 1, the dashed line to PCODE = 2, and the dotted line to PCODE = 3.

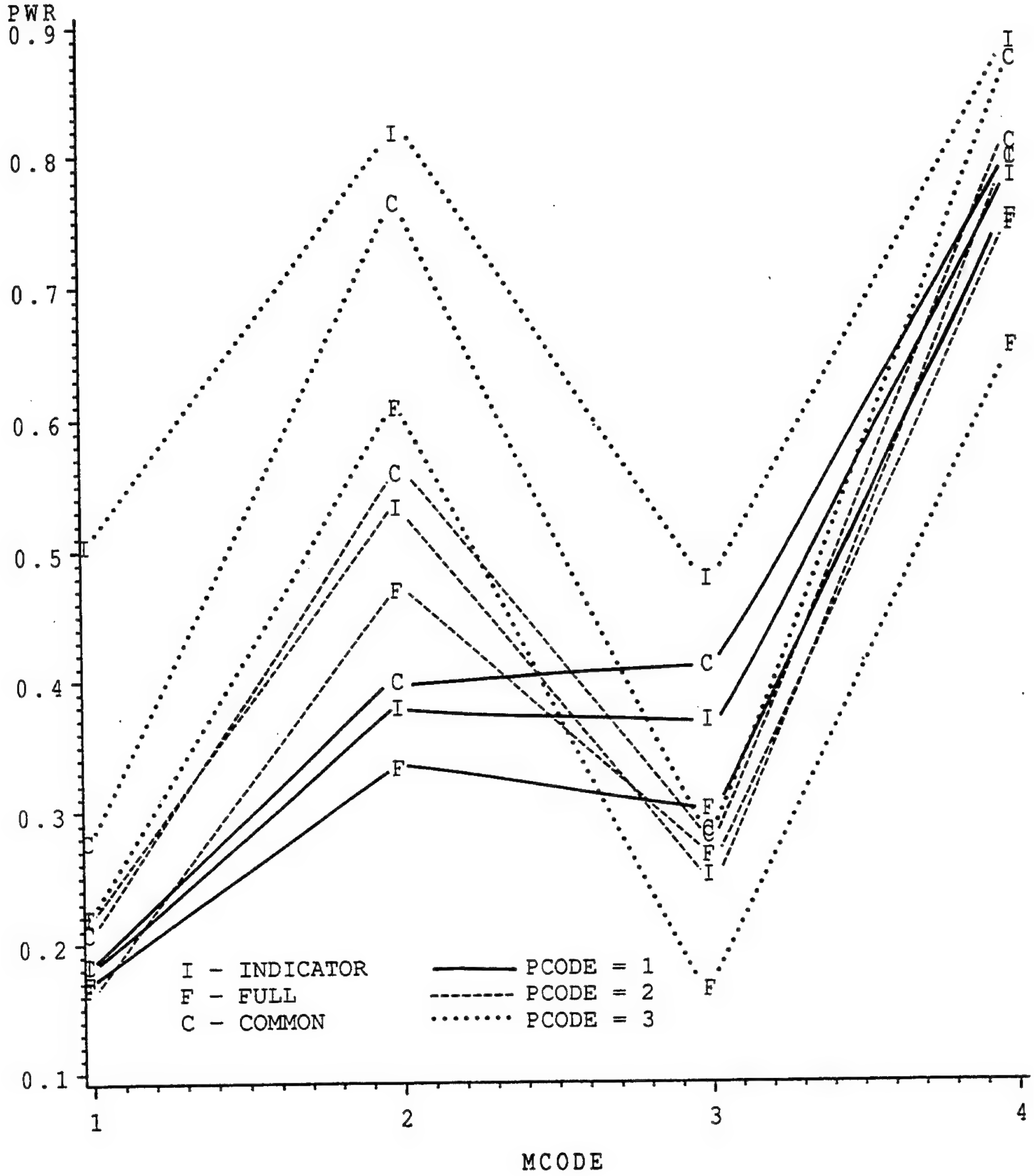
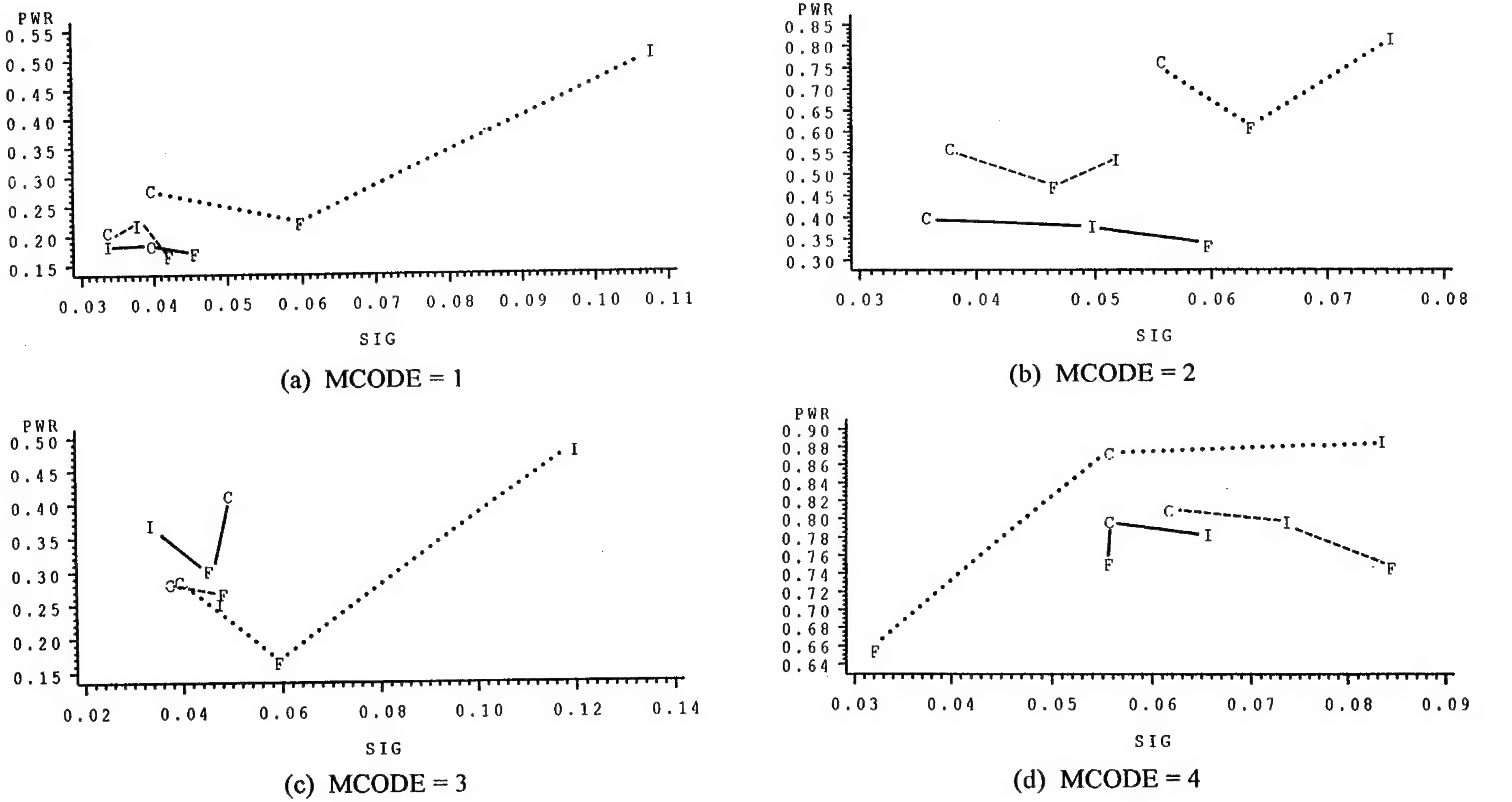


Figure 11. Power estimates when approximately 10% of the data is missing for each of the three algorithms when data have two binary and two continuous variates, possibly dependent. For population π_1, each possible combination of the binary part occurs with probability 1/4. The conditional distribution of the continuous part is MVN($\mathbf{0}$, Σ_1), where Σ_1 is a 2x2 matrix with diagonal elements of one and off-diagonal elements of 0.5, within each multinomial cell. For population π_2, the conditional covariance matrix for the continuous part is Σ_2, where Σ_2 has ones on the diagonal and off-diagonal elements of -0.5. Several distributions for the discrete part, and several choices of mean vectors are used, as defined in Table 2.



(a) MCODE = 1

(b) MCODE = 2

(c) MCODE = 3

(d) MCODE = 4

Figure 12. Plots of estimated power versus observed significance for each mean configuration in our second study when 10% of the data is missing. Symbol used is I for INDICATOR algorithm, F for FULL, and C for COMMON. The solid line corresponds to PCODE = 1, the dashed line to PCODE = 2, and the dotted line to PCODE = 3.

Outlier Tests with Multiple Stations

H. L. Gray, Wayne A. Woodward and Zeynep T. Yücel
Southern Methodist University

August 25, 1995

Abstract

Some techniques are discussed for dealing with the problem of distinguishing between earthquakes and explosions when data are available at more than one station. A simulation study, in which the performance of these techniques are compared, is presented.

1 Introduction

We consider the problem of observing seismic events for the purpose of distinguishing between earthquakes and explosions. Baek, et al. (1994) treat this as an outlier problem which is to determine whether a new and possibly a suspicious event should be classified as an earthquake, given a training sample of data on earthquakes. In that paper, it was assumed that data on several variables were available at a given station. It was also assumed that the variables might be either continuous, discrete or a mixture of both types.

In this report, we address the issue of outlier testing when data are at more than one station. In particular, we suppose that there are p feature variables observed at each of m stations. A fundamental problem is how to utilize the information from multiple stations in a test for outliers.

1.1 Test 1: Full Vector Approach

One technique for outlier detection in the multi–station case is to consider the p features at each of m stations as a single vector consisting of mp variables. That is, the observation vector for the i th event in the training sample is an $mp \times 1$ vector of the form

$$\mathbf{X}_i = (X_{11i}, X_{12i}, \ldots, X_{1mi}, X_{21i}, X_{22i}, \ldots, X_{2mi} \ldots, X_{p1i}, X_{p2i}, \ldots, X_{pmi})' ,$$

$i = 1, \ldots, n$, where X_{jki} indicates the j th feature measured at the k th station for the i th earthquake. A new observation to be tested as an outlier has then a similarly configured $mp \times 1$ vector of the form

$$\mathbf{X}_{n+1} = (X_{11,n+1}, X_{12,n+1}, \ldots X_{1m,n+1}, \ldots, X_{p1,n+1}, X_{p2,n+1}, \ldots, X_{pm,n+1})' .$$

We consider the training sample $\{\mathbf{X}_i\}_{i=1}^{n}$ to be from the density function $f(.; \boldsymbol{\mu}_1, \Sigma)$, where

$$f(\mathbf{X}; \boldsymbol{\mu}_1, \Sigma) = (2\,\pi)^{-\frac{mp}{2}}\,|\Sigma|^{-\frac{1}{2}}\, \exp\left\{-\frac{1}{2}\,(\mathbf{X} - \boldsymbol{\mu}_1)'\,\Sigma^{-1}\,(\mathbf{X} - \boldsymbol{\mu}_1)\right\} ,$$

i.e., we are assuming in this report that the feature variables have a multivariate normal distribution. Similarly, the new $\mathbf{X}_{n+1}$ is assumed to have probability density $f(.; \boldsymbol{\mu}_2, \Sigma)$. Baek, et al. (1994) classify $\mathbf{X}_{n+1}$ by testing the hypotheses

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$
$$H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2.$$

The likelihood of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \mathbf{X}_{n+1}$ is given by

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{X}_1, \ldots, \mathbf{X}_{n+1}) &= L(\boldsymbol{\theta}; \mathbf{X}_1, \ldots, \mathbf{X}_n)\,(2\,\pi)^{-\frac{mp}{2}}\,|\Sigma|^{-\frac{1}{2}} \\ &\quad \exp\left\{-\frac{1}{2}\,(\mathbf{X}_{n+1} - \boldsymbol{\mu}_2)'\,\Sigma^{-1}\,(\mathbf{X}_{n+1} - \boldsymbol{\mu}_2)\right\} , \end{aligned}$$

where $L(\boldsymbol{\theta}; \mathbf{X}_1, \ldots, \mathbf{X}_n) = \prod_{i=1}^{n} f(\mathbf{X}_i; \boldsymbol{\mu}_1, \Sigma)$, the likelihood of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, and $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$. The generalized likelihood ratio is therefore defined by

$$\begin{aligned} \lambda &= \frac{\sup_{\{\boldsymbol{\theta} \in \Omega_0\}} L(\boldsymbol{\theta}; \mathbf{X}_1, \ldots, \mathbf{X}_{n+1})}{\sup_{\{\boldsymbol{\theta} \in \Omega\}} L(\boldsymbol{\theta}; \mathbf{X}_1, \ldots, \mathbf{X}_{n+1})} \\ &= \frac{L(\hat{\boldsymbol{\theta}}_0; \mathbf{X}_1, \ldots, \mathbf{X}_{n+1})}{L(\hat{\boldsymbol{\theta}}; \mathbf{X}_1, \ldots, \mathbf{X}_{n+1})}, \end{aligned} \tag{1}$$

where $\hat{\boldsymbol{\theta}}_0$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ under the restriction that H_0 is true, and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\Sigma})$ where $\hat{\boldsymbol{\mu}}_1$ and $\hat{\Sigma}$ are the MLE's of $\boldsymbol{\mu}_1$ and Σ based on $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ and $\hat{\boldsymbol{\mu}}_2 = \mathbf{X}_{n+1}$. It intuitively follows that small values of λ provide evidence against H_0, and thus the generalized likelihood ratio test is

$$\text{reject} \quad H_0 \quad \text{if} \quad \lambda \leq \lambda(\alpha), \tag{2}$$

where $\lambda(\alpha)$ is chosen to provide a size α test. In the setting considered by Baek, et al. (1994), i.e. the data are a mixture of continuous and discrete types, the distribution of λ is unknown: and bootstrap techniques were applied in order to ascertain this distribution. Baek, et al. (1994) point out that when all classification variables are continuous and have a multivariate normal distribution, the setting considered here. the distribution of λ is known; and the critical values can be found based on the F–distribution. In particular, in the current case of mp variables and a training sample of size n, $\lambda(\alpha)$ is given by

$$\lambda(\alpha) = \left(1 + \frac{m\,p\,F_\alpha}{n - m\,p}\right)^{-\frac{n+1}{2}}, \tag{3}$$

where F_α is the $(1-\alpha)$ th percentile of the F–distribution with mp and $n - mp$ degrees of freedom. In the multivariate normal case, calculating the critical value $\lambda(\alpha)$ by using the bootstrap procedure or from (3) produced very similar results[1]. More details concerning the derivation of (3) are given in Fisk (1995).

In the full vector approach, no attempt is made to account for the fact that the same p variables are being measured at the m stations. It should be noted that this solution strategy actually does not require the same p variables to be observed at each of the m stations.

1.2 Test 2: Minimum Variance Weighting

A second method that will be examined here is the combining of features across stations by using minimum variance weighting. This procedure is

[1] These simulations results will not be given here.

designed to reduce the dimensionality of the problem by taking advantage of the correlation structure between stations. We construct a new feature Y_j associated with feature j. The new feature is the linear combination of feature j at each of the m stations, i.e.,

$$Y_j = \sum_{k=1}^{m} \omega_{jk} X_{jk} , \tag{4}$$

which minimizes the variance of Y_j subject to the constraint that the weights sum to one. Theoretically, the weights are given by

$$\omega_j = \frac{\Sigma_j^{-1} \boldsymbol{\beta}}{\boldsymbol{\beta}' \Sigma_j^{-1} \boldsymbol{\beta}} , \tag{5}$$

where $\boldsymbol{\beta} = (1, 1, \ldots, 1)'$ and Σ_j is the covariance matrix of $X_{j1}, \ldots, X_{jm}$. In practice, Σ_j will not be known, and will be estimated by the usual sample covariance matrix based on events $i = 1, \ldots, n$. This weighting is not based on the assumption that the means for the j th feature are constant across stations, but rather is combining the data across stations to create a new feature. This procedure creates a new p–dimensional vector $\mathbf{Y}_i = (Y_{1i}, \ldots, Y_{pi})'$. $i = 1, \ldots, n$. The $(n+1)$ st event, which is to be classified as a possible outlier, is weighted by using the same weights, i.e.,

$$Y_{j,n+1} = \sum_{k=1}^{m} \omega_{jk} X_{jk,n+1} . \tag{6}$$

This weighting reduces the dimension from mp variables to p variables. The outlier detection is then based on a likelihood ratio as before but calculated using only the p new variables. It should be noted that although the weights are stochastic and depend on the data, for feature j the same m station weights, ω_{jk}, $k = 1, \ldots, m$, are used for each of the events. Thus, the resulting new p features will be approximately normally distributed random variables.

1.3 Test 3: Separate Tests Based on Each Station

It is possible for the test based on all stations to fail to declare an event to be an outlier while the test based on one of the individual stations finds the event to be an outlier. It seems plausible that a very noisy station could result in the multi–station test losing power compared to that associated with an individual "good" station[2]. The question is whether multi–station tests are more powerful than simple use of individual station–based tests. An obvious strategy for using station information at m stations is to declare an event

[2] The identification of "good" stations for use in the test is under investigation and will be the subject of the future report.

to be an outlier if any of the individual station–based tests finds the event to be an outlier. This has the apparent advantage that if, for example, there are two stations with one of the stations quite noisy, then the test has not been penalized for inclusion of the noisy station, as may be the case with the other two tests considered. However, if each of the tests at the m stations is run at the $\alpha = 0.05$ level, the result of such a procedure will be a test with significance level larger than α, and according to Bonferroni's inequality, the overall significance level is less than or equal to $m\alpha$. Thus, in order to assure that the overall significance level is no more than α, each individual station–based test should be run at level α/m. In the example of two stations, one good and one noisy, this procedure also provides a penalty for incorporating the noisy station since it requires implementation of a smaller significance level, $\alpha = 0.025$, for each test. Thus, if the second station is of no usefulness, then the result of the procedure is to reduce the power over that of identifying the useful station and applying the test using only that station. Note that the weighting done here is optimal (in the estimation sense) for the case in which the variables at different stations are independent, which is not the case here.

In the following section, we present some simulation results in which we compare the power of the three outlier testing procedures described here in order to examine the conditions for which a particular test is favored over the others.

2 Simulation Examination of the Tests

2.1 Two Stations and One Variable

In this subsection, we consider the 2–dimensional case of two stations and one variable measured at each station. For the population of earthquakes, we assume that

$$\mathbf{X}_i = (X_{11i}, X_{12i})' \sim MVN(\boldsymbol{\mu}_1, \Sigma) , \qquad (7)$$

where

$$\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12})' \qquad \text{with} \qquad \mu_{11} = \mu_{12} = 0$$

and

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \qquad \text{for} \qquad \rho = -0.25, 0.0, 0.25, 0.50, 0.75 .$$

In Table 1, we present estimated power of the full vector, minimum variance weighting, and separate station–based tests when the potential outlier is from

$$\mathbf{X}_{n+1} \sim MVN(\boldsymbol{\mu}_1^{(A)}, \Sigma) , \qquad (8)$$

where $\boldsymbol{\mu}_1^{(A)} = (\mu_{11}^{(A)}, \mu_{12}^{(A)})'$. In each case, a sample of size $n = 50$ is generated from the earthquake distribution, i.e., $MVN((0,0)', \Sigma)$, and a single outlier is generated from the outlier distribution, i.e., $MVN((\mu_{11}^{(A)}, \mu_{12}^{(A)})', \Sigma)$ for a variety of values of $\mu_{11}^{(A)}$ and $\mu_{12}^{(A)}$. The test determines whether this single observation is classified as an outlier. The entire procedure is repeated 1000 times, and the power estimates given in the table are the proportion of times that the single observation from $MVN((\mu_{11}^{(A)}, \mu_{12}^{(A)})', \Sigma)$ was called as an outlier. All tests presented in this report were run at the $\alpha = 0.05$ overall significance level.

We first focus on the full vector test, Test 1. There we can see in general, as $\mu_{11}^{(A)}$ and $\mu_{12}^{(A)}$ become further removed from the null values of 0, the power of Test 1 increases as would be expected. However, some of the results in the table may seem nonintuitive at first glance. For example, when $\rho = 0.75$ the power associated with the alternative $\mu_{11}^{(A)} = 2$ and $\mu_{12}^{(A)} = 0$ is 0.738 which is much higher than for the case $\rho = 0$ in which the power is 0.377. In some respects this seems to be an unusual result since intiutively it would seem that highly correlated variables (stations) would tend to be providing redundant information and hence might be expected to yield lower power than in the case in which the correlation is smaller. However, it should be pointed out that while increased correlation reduces information in estimation, it may dramatically increase information for purposes of outlier detection. Thus, it is important to note that the shape of the bivariate distribution plays a major role in determining this power, i.e., in determining what types of values appear to be outliers. In Figure 1, we show contour plots of the bivariate distributions assumed under the null hypothesis for the values of ρ considered. There it can be seen that observations around $\mu_{11}^{(A)} = 2$ and $\mu_{12}^{(A)} = 0$ for station 1 and station 2, respectively, are much more unlikely when $\rho = 0.75$ than for lower values of ρ. Interpretation of other powers shown in Table 1 is aided by examination of Figure 1.

It can be seen from Table 1 that the minimum variance weighting test, Test 2, results are sometimes comparable and in some instances superior to those obtained using Test 1, the full vector approach. However, it is also noted that in some cases these powers can be much worse than those obtained using Test 1. In order to understand this phenomenon, consider again the case in which $\mu_{11}^{(A)} = 2$ and $\mu_{12}^{(A)} = 0$ with $\rho = 0.75$. The effect of the minimum variance weighting is to produce new feature Y_1 calculated as $Y_{1i} = \omega_{11} X_{11i} + \omega_{12} X_{12i}$, $i = 1, \ldots, n$. In this case, the weights will both be approximately equal so that the mean of Y_1 will be about 1, and in Table 1 it is seen that the power of Test 2 is only 0.185 when $\rho = 0.75$ as compared to 0.738 using Test 1. From Table 1 we see that for $\mu_{11}^{(A)}$ relatively close to $\mu_{12}^{(A)}$ Test 2 tends to have higher power than Test 1. It is clear from Table 1 that if Test 2 care must be taken. In the case of positive correlations between stations Test 2 will tend to perform poorly when the potential outliers are not consistent with this correlation structure. The fact that $(2, 0)'$ is not consistent with the correlation structure can be measured using the Mahalanobis distance, defined as

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^{(A)})' \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^{(A)}), \tag{9}$$

between the null and the alternative populations. Thus, although the points $(2, 0)'$ and $(\sqrt{2}, \sqrt{2})'$ have the same squared Euclidean distance, 4, from the null mean of $(0, 0)'$, the Mahalanobis distance for $(2, 0)'$ is 4 times as large as that for $(\sqrt{2}, \sqrt{2})'$ whenever $\rho = 0.75$. Thus, when the correlations between stations are positive, a large Mahalanobis distance, compared to the range of Mahalanobis distances possible for a given Euclidean distance, should serve as an indication that Test 2 should not be used. The procedure we used in the simulations for assessing whether minimum variance weighting should be used in a situation in which the correlations between stations are positive is given below:

1. Calculate the Euclidean distance $d_E = (\boldsymbol{\mu}_1 - X_{n+1})' (\boldsymbol{\mu}_1 - X_{n+1})$ and the Mahalanobis distance $d_M = (\boldsymbol{\mu}_1 - X_{n+1})' \Sigma^{-1} (\boldsymbol{\mu}_1 - X_{n+1})$ between the null mean and the potential outlier, (for $\boldsymbol{\mu}_1$ and Σ, use sample values calculated from the training sample).

2. Calculate the minimum, $d_M^{(\min)}$, and first quartile, $d_M^{(Q_1)}$, of all possible Mahalanobis distances associated with means separated by a Euclidean distance d_E.

3. Whenever $d_M / d_M^{(\min)} < 2$ and $d_M < d_M^{(Q_1)}$, minimum variance weighting is appropriate. Otherwise, the full vector approach is recommended.

In Table 1, we give the power of a "combined test". Test 4, which, for a given training sample and potential outlier, uses Test 1 or Test 2 as indicated by the Mahalanobis distance criterion mentioned above. The Mahalanobis distance criterion is specifically designed for the case in which there is positive correlation. It can be seen that in cases in which $\rho \geq 0.25$, Test 4 often performed better than the other three tests and always had close to the highest power. Test 4 did not perform as well for $\rho = 0$. However, this test is designed for the case in which the correlation is positive, and the results for $\rho = 0$ are included only for comparison. It should be noted that, in practice, the decision concerning whether to use the Mahalanobis check will be based on the correlations calculated from the training sample data; and some rules should be obtained for deciding how large a sample correlation should be before the Mahalanobis check is performed. It should be noted that Test 4 tends to have slightly larger significance level, (i.e., power at the alternative $(0, 0)'$) than the nominal level of $\alpha = 0.05$.

In Table 1, we also give the power results associated with Test 3. In variable at each of two stations, the associated univariate test is run for each of the two stations at $\alpha = 0.025$, and an outlier is said to be detected if either of the two univariate tests determines the event to be an outlier. As indicated earlier, this assures that the overall test has significance level no more than $\alpha = 0.05$. In the table we see that the power is competitive with Tests 1 and 4 except in the case in which the correlation is large ($\rho \geq 0.50$) and one of $\mu_{11}^{(A)}$ or $\mu_{12}^{(A)}$ is close to the null value of zero.

In Table 2, we present simulation results for a case in which the station variances are not equal. In particular, we consider the case in which

$$\Sigma = \begin{pmatrix} 1 & 2\rho \\ 2\rho & 4 \end{pmatrix} \quad \text{with} \quad \rho = -0.25, 0.0, 0.25, 0.50, 0.75 \; . \qquad (10)$$

Figure 2 shows the contour plots of the bivariate distributions assumed under the null hypothesis for the values of ρ considered. In this setting, we observe that Test 1 and Test 2 behave similarly in the sense that, in general, as $\mu_{11}^{(A)}$ and $\mu_{12}^{(A)}$ become further apart from the null values of 0, and as correlation between stations increases, the powers of these tests increase. As in Table 1, while Test 2 results are often comparable or even superior to Test 1. for some cases. Test 2 results are much worse than those for Test 1. This seems particularly true for values of $\mu_{11}^{(A)} = 0$ and $\mu_{12}^{(A)} > 0$ which, as can be seen in Figure 2. are values that do not correspond to the correlation structure. Using the Mahalanobis distance criterion as in Table 1 we obtained Test 4 which still has substantially lower power than Test 1 in the cases where Test 2 power is much less than Test 1. As in Table 1 we see that Test 1 has power which is always competitive with the best shown in the table whereas each of the other tests can in some instances have substantially lower power than Test 1.

We also have examined the use of outlier tests using model parameters obtained from actual seismic data. Data are available on 36 earthquakes and 70 explosions for the logarithm of Pn/Lg(6–8Hz) at two stations, KNB and MNV. Letting KNB be station 1 and MNV station 2, the sample mean vector and covariance matrix for this training sample of earthquakes is given by

$$\bar{\mathbf{X}}_1 = (\bar{X}_{11}, \bar{X}_{12})' = (-0.743, -1.672)' \quad \text{and} \quad \hat{\Sigma}_1 = \begin{pmatrix} 0.235 & 0.046 \\ 0.046 & 0.173 \end{pmatrix} , \quad (11)$$

while the corresponding quantities for the training sample of explosions are

$$\bar{\mathbf{X}}_2 = (0.361, -0.010)' \quad \text{and} \quad \hat{\Sigma}_2 = \begin{pmatrix} 0.196 & 0.053 \\ 0.053 & 0.168 \end{pmatrix} , \qquad (12)$$

The estimated correlation between stations is found to be 0.23 earthquakes; and 0.29 for the training samples of explosions. The contour plots for the bivariate normals with means and covariances set equal to those observed for the two training samples are given in Figure 3. There it can be seen that there is a substantial separation between the two populations; and it would be expected that the outlier test should be able to detect explosions as outliers related to the earthquake population. In order to examine this, we performed the outlier tests using samples of length 50 simulated from the model with $\boldsymbol{\mu}_1 = \bar{\mathbf{X}}_1$ and $\Sigma_1 = \hat{\Sigma}_1$ with outliers simulated from the bivariate normal model with $\boldsymbol{\mu}_2 = \bar{\mathbf{X}}_2$ and $\Sigma_2 = \hat{\Sigma}_2$. In this setting, an outlier was detected 95.3%, 95.6%, 97.5% and 96.2% of the time by using the separate station-based, full vector, minimum variance weighting and combined full vector-minimum variance weighting tests, respectively.

2.2 Two Stations and Two Variables

In this subsection, we briefly consider the case in which there are two feature variables measured at each of two stations.

To compare the performances of Test 1, Test 2, Test 3 and Test 4 of section 1 under this setting, we have carried out some simulations on data generated from multivariate normal distribution with various mean vectors and covariance matrices. From the simulation study, we have observed that, as in the case of two stations-one variable, the estimated powers of Test 2 are sometimes comparable and in some cases superior to those of Test 1: and there are also some cases in which powers of Test 2 are much worse than those of Test 1. The combined test, Test 4, performs fairly well, but it is clear that the Mahalanobis decision rule on page 8 may not be optimal for a wide range of parameter configurations.

We also have examined the powers of the outlier tests of section 1 using model parameters obtained from actual seismic data. In the simulations, we assume the observations $\mathbf{X}_i = (X_{11i}, X_{12i}, X_{21i}, X_{22i})'$, $i = 1, \ldots, n$ are from $MVN(\boldsymbol{\mu}_1, \Sigma)$. Data are available on 36 earthquakes and 70 explosions for the logarithm of Pn/Lg ratio in both the 4–6Hz and 6–8Hz frequency bands. We let station 1 denote KNB and station 2 denote MNV, and we take the log Pn/Lg ratio in the 6–8Hz and 4–6Hz frequency bands as features 1 and 2, respectively. The sample mean vector and covariance matrix for this training sample of earthquakes is given by

$$\bar{\mathbf{X}}_1 = (\bar{X}_{11}, \bar{X}_{12}, \bar{X}_{21}, \bar{X}_{22})' = (-0.712, -0.992, -1.657, -1.829)' \qquad (13)$$

and

$$\hat{\Sigma}_1 = \begin{pmatrix} 0.242 & 0.202 & 0.052 & 0.044 \\ 0.202 & 0.275 & 0.052 & 0.075 \\ 0.052 & 0.052 & 0.177 & 0.086 \\ 0.044 & 0.075 & 0.086 & 0.227 \end{pmatrix},$$

while the corresponding quantities for the training sample of explosions are

$$\bar{\mathbf{X}}_2 = (0.362, -0.155, -0.032, -0.453)' \tag{14}$$

and

$$\hat{\Sigma}_2 = \begin{pmatrix} 0.197 & 0.138 & 0.054 & 0.012 \\ 0.138 & 0.195 & 0.050 & -0.009 \\ 0.054 & 0.050 & 0.180 & 0.093 \\ 0.012 & -0.009 & 0.093 & 0.207 \end{pmatrix} .$$

Let $\hat{\rho}^{(k_1,k_2)}_{j_1,j_2}$ denote the estimated correlation between features j_1 and j_2 at stations k_1 and k_2. These quantities for the training sample of earthquakes are

$$\hat{\rho}^{(1,2)}_{1,1} = 0.3 \ , \ \hat{\rho}^{(1,2)}_{2,2} = 0.25 \ , \ \hat{\rho}^{(1,1)}_{1,2} = 0.43 \ , \ \hat{\rho}^{(2,2)}_{1,2} = 0.78 \ , \ \hat{\rho}^{(1,2)}_{1,2} = 0.19 \ . \tag{15}$$

The corresponding quantities for the training sample of explosions are

$$\hat{\rho}^{(1,2)}_{1,1} = -0.05 \ , \ \hat{\rho}^{(1,2)}_{2,2} = 0.28 \ , \ \hat{\rho}^{(1,1)}_{1,2} = 0.48 \ , \ \hat{\rho}^{(2,2)}_{1,2} = 0.70 \ , \ \hat{\rho}^{(1,2)}_{1,2} = 0.06 \ . \tag{16}$$

Figure 4 displays the contour plots for the bivariate normals with means and covariances set equal to those observed for the two training samples for each of the feature variables at two stations. There it can be seen that for each of the feature variables, there is a reasonable separation between the two populations; and it would be expected that the outlier test should be able to detect explosions as outliers related to the earthquake population. In order to examine this, we performed the outlier tests using samples of length 50 simulated from the model with $\boldsymbol{\mu}_1 = \bar{\mathbf{X}}_1$ and $\Sigma_1 = \hat{\Sigma}_1$ with outliers simulated from the multivariate normal model with $\boldsymbol{\mu}_2 = \bar{\mathbf{X}}_2$ and $\Sigma_2 = \hat{\Sigma}_2$. In this setting, an outlier was detected 81.1%, 89.1%, 94.1% and 93.9% of the time by using the separate station-based, full vector, minimum variance weighting and combined full vector-minimum variance weighting tests, respectively. We have also examined the outlier tests of section 1 to determine their success rates in classifying the 70 events in the training sample of explosions as outliers. The success rates[3] for Test 1, Test 2, Test 3 and the combined test are 92.9%, 92.9%, 77.1% and 91.4% respectively. Test 1 and Test 2 agreed in classifying these 70 events except for two cases. In one of these two cases Test 1 does not classify the event considered as an outlier, and Mahalanobis distance criterion favors Test 1. In the second case, Test 2 does not classify the event considered as an outlier, and Mahalanobis distance criterion favors Test 2. The Mahalanobis distance criterion favors Test 1 51.4% of the time, and favors Test 2 48.6% of the time.

[3] The test classifies the event from the training sample of explosions as outlier.

3 Tables and Figures

- **Table 1.** Estimated powers of the outlier tests considered under the setting:

  $$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \text{ with } \rho = -0.25, 0.0, 0.25, 0.50, 0.75$$

  and $\mu_{11}^{(A)}, \mu_{12}^{(A)} = 0, 1, 2, 3, 4$. (The values in the parantheses are the number of times that full vector approach and minimum variance weighting approach performed, respectively, for the combined test.)

- **Table 2.** Estimated powers of the outlier tests considered under the setting:

  $$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 2\rho \\ 2\rho & 4 \end{pmatrix} \text{ with } \rho = -0.25, 0.0, 0.25, 0.50, 0.75$$

  and $\mu_{11}^{(A)}, \mu_{12}^{(A)} = 0, 1, 2, 3, 4$. (The values in the parantheses are the number of times that full vector approach and minimum variance weighting approach performed, respectively, for the combined test.)

- **Figure 1.** Contour plots of bivariate normal distributions from

  $$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \text{ with } \rho = -0.25, 0.0. 0.25, 0.50, 0.75 .$$

- **Figure 2.** Contour plots of bivariate normal distributions from

  $$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 2\rho \\ 2\rho & 4 \end{pmatrix} \text{ with } \rho = -0.25, 0.0. 0.25, 0.50. 0.75 .$$

- **Figure 3.** (Two stations–one variable case) Contour plots of bivariate normal distributions with means and covariances calculated from actual seismic data:

  $$\bar{\mathbf{X}}_1 = \begin{pmatrix} -0.743 \\ -1.672 \end{pmatrix}, \hat{\Sigma}_1 = \begin{pmatrix} 0.235 & 0.046 \\ 0.046 & 0.173 \end{pmatrix},$$

  $$\bar{\mathbf{X}}_2 = \begin{pmatrix} 0.361 \\ -0.010 \end{pmatrix}, \hat{\Sigma}_2 = \begin{pmatrix} 0.196 & 0.053 \\ 0.053 & 0.168 \end{pmatrix}.$$

- **Figure 4.** (Two stations–two variables case) Contour plots of bivariate normal distributions of each variable with means and covariances calculated from actual seismic data:

  For variable 1;

$$\bar{\mathbf{X}}_1 = \begin{pmatrix} -0.712 \\ -1.657 \end{pmatrix} \ , \ \hat{\Sigma}_1 = \begin{pmatrix} 0.242 & 0.052 \\ 0.052 & 0.177 \end{pmatrix} \ ,$$

$$\bar{\mathbf{X}}_2 = \begin{pmatrix} 0.362 \\ -0.032 \end{pmatrix} \ , \ \hat{\Sigma}_2 = \begin{pmatrix} 0.197 & 0.054 \\ 0.054 & 0.180 \end{pmatrix} \ .$$

  For variable 2;

$$\bar{\mathbf{X}}_1 = \begin{pmatrix} -0.992 \\ -1.829 \end{pmatrix} \ , \ \hat{\Sigma}_1 = \begin{pmatrix} 0.275 & 0.075 \\ 0.075 & 0.227 \end{pmatrix} \ ,$$

$$\bar{\mathbf{X}}_2 = \begin{pmatrix} -0.155 \\ -0.453 \end{pmatrix} \ , \ \hat{\Sigma}_2 = \begin{pmatrix} 0.195 & -0.009 \\ -0.009 & 0.207 \end{pmatrix} \ .$$

Table 1.[4]

$\rho = -0.25$	Test	$\mu_{11}^{(A)}$				
$\mu_{12}^{(A)}$		0	1	2	3	4
0	1	0.044	0.129	0.389	0.751	0.949
	2	0.054	0.137	0.347	0.648	0.877
	3	0.041	0.121	0.367	0.740	0.945
1	1	0.122	0.250	0.560	0.856	0.970
	2	0.141	0.348	0.641	0.884	0.967
	3	0.117	0.197	0.430	0.769	0.952
2	1	0.400	0.570	0.793	0.947	0.989
	2	0.360	0.654	0.891	0.973	0.996
	3	0.390	0.450	0.629	0.857	0.966
3	1	0.749	0.848	0.943	0.985	0.999
	2	0.650	0.882	0.974	0.996	1.000
	3	0.745	0.777	0.863	0.948	0.986
4	1	0.951	0.975	0.993	1.000	1.000
	2	0.871	0.970	0.995	1.000	1.000
	3	0.948	0.959	0.983	0.996	1.000

[4] One table for each ρ considered.

Table 1. continues:

$\rho = 0.0$	Test	$\mu_{11}^{(A)}$				
$\mu_{12}^{(A)}$		0	1	2	3	4
0	1	0.044	0.122	0.377	0.723	0.936
	2	0.056	0.119	0.282	0.525	0.766
	3	0.047	0.128	0.376	0.748	0.948
	4	0.043 (751, 249)	0.122 (743. 257)	0.346 (752, 248)	0.671 (726, 274)	0.875 (725, 275)
1	1	0.115	0.203	0.446	0.773	0.946
	2	0.123	0.274	0.520	0.782	0.926
	3	0.116	0.192	0.418	0.758	0.947
	4	0.115 (763, 237)	0.217 (764. 236)	0.462 (746, 254)	0.770 (748, 252)	0.931 (733. 267)
2	1	0.381	0.455	0.665	0.878	0.967
	2	0.284	0.529	0.789	0.930	0.975
	3	0.393	0.441	0.605	0.838	0.962
	4	0.369 (765, 235)	0.466 (756. 244)	0.701 (760, 240)	0.896 (754. 246)	0.969 (741, 259)
3	1	0.726	0.760	0.877	0.957	0.988
	2	0.534	0.782	0.926	0.979	0.996
	3	0.749	0.770	0.841	0.932	0.980
	4	0.675 (754, 246)	0.761 (751. 249)	0.890 (752. 248)	0.965 (764. 236)	0.989 (747, 253)
4	1	0.932	0.946	0.969	0.987	0.999
	2	0.766	0.917	0.977	0.997	1.000
	3	0.939	0.945	0.964	0.983	0.992
	4	0.871 (753, 247)	0.931 (756. 244)	0.972 (751, 249)	0.990 (758. 242)	0.999 (763, 237)

Table 1. continues:

$\rho = 0.25$	Test	$\mu_{11}^{(A)}$				
$\mu_{12}^{(A)}$		0	1	2	3	4
0	1	0.044	0.140	0.399	0.756	0.957
	2	0.056	0.105	0.245	0.458	0.678
	3	0.045	0.128	0.378	0.750	0.950
	4	0.055 (696, 304)	0.157 (717, 283)	0.415 (764, 236)	0.748 (824, 176)	0.939 (855, 145)
1	1	0.122	0.171	0.389	0.727	0.936
	2	0.106	0.238	0.436	0.690	0.845
	3	0.117	0.186	0.406	0.748	0.945
	4	0.145 (735, 265)	0.226 (577, 423)	0.453 (535, 465)	0.765 (595, 405)	0.927 (660, 340)
2	1	0.400	0.397	0.560	0.802	0.949
	2	0.245	0.442	0.691	0.868	0.953
	3	0.383	0.414	0.562	0.805	0.951
	4	0.417 (789, 211)	0.454 (554, 446)	0.671 (398, 602)	0.864 (391. 609)	0.963 (464, 536)
3	1	0.749	0.723	0.793	0.906	0.970
	2	0.452	0.674	0.868	0.958	0.984
	3	0.736	0.745	0.800	0.900	0.968
	4	0.741 (826. 174)	0.754 (632, 368)	0.853 (403, 597)	0.941 (294. 706)	0.983 (313, 687)
4	1	0.951	0.934	0.943	0.974	0.989
	2	0.674	0.857	0.952	0.983	0.997
	3	0.942	0.939	0.950	0.970	0.986
	4	0.938 (851, 149)	0.934 (700, 300)	0.959 (492, 508)	0.982 (311. 689)	0.996 (256, 744)

Table 1. continues:

$\rho = 0.50$	Test	$\mu_{11}^{(A)}$				
$\mu_{12}^{(A)}$		0	1	2	3	4
0	1	0.044	0.161	0.481	0.865	0.992
	2	0.056	0.098	0.213	0.410	0.588
	3	0.046	0.128	0.379	0.752	0.954
	4	0.063 (607, 393)	0.183 (683, 317)	0.501 (825, 175)	0.863 (903, 97)	0.987 (936, 64)
1	1	0.149	0.151	0.377	0.742	0.961
	2	0.096	0.214	0.378	0.603	0.783
	3	0.123	0.180	0.393	0.741	0.946
	4	0.176 (698, 302)	0.222 (464, 536)	0.446 (466, 534)	0.772 (620, 380)	0.953 (717, 283)
2	1	0.489	0.371	0.474	0.733	0.936
	2	0.221	0.388	0.604	0.798	0.909
	3	0.374	0.391	0.518	0.775	0.947
	4	0.499 (807, 193)	0.438 (488, 512)	0.605 (231, 769)	0.822 (251, 749)	0.948 (407, 593)
3	1	0.842	0.736	0.740	0.851	0.954
	2	0.404	0.594	0.794	0.929	0.969
	3	0.737	0.735	0.774	0.873	0.962
	4	0.838 (892, 108)	0.758 (623, 377)	0.822 (258, 742)	0.923 (96, 904)	0.972 (136, 864)
4	1	0.982	0.951	0.935	0.950	0.976
	2	0.591	0.783	0.917	0.969	0.990
	3	0.940	0.935	0.938	0.957	0.978
	4	0.980 (931, 69)	0.946 (752, 248)	0.951 (413, 587)	0.967 (136, 864)	0.990 (51, 949)

Table 1. continues:

$\rho = 0.75$	Test	$\mu_{11}^{(A)}$				
$\mu_{12}^{(A)}$		0	1	2	3	4
0	1	0.044	0.244	0.738	0.989	1.000
	2	0.053	0.094	0.185	0.370	0.537
	3	0.044	0.127	0.378	0.753	0.956
	4	0.066 (513, 487)	0.258 (705.295)	0.730 (882, 118)	0.982 (956, 44)	0.999 (978, 22)
1	1	0.238	0.135	0.425	0.880	0.999
	2	0.099	0.188	0.348	0.545	0.701
	3	0.113	0.154	0.370	0.740	0.946
	4	0.259 (693, 307)	0.204 (364, 636)	0.485 (459, 541)	0.865 (683, 317)	0.994 (843, 157)
2	1	0.728	0.429	0.413	0.729	0.979
	2	0.200	0.338	0.538	0.741	0.862
	3	0.365	0.370	0.473	0.749	0.945
	4	0.720 (869, 131)	0.461 (466, 534)	0.544 (130, 870)	0.804 (224, 776)	0.961 (420, 580)
3	1	0.980	0.855	0.733	0.788	0.936
	2	0.355	0.531	0.728	0.874	0.950
	3	0.740	0.732	0.747	0.837	0.948
	4	0.974 (947, 53)	0.850 (697, 303)	0.789 (212, 788)	0.876 (31.969)	0.964 (69, 931)
4	1	1.000	0.992	0.952	0.939	0.961
	2	0.543	0.693	0.862	0.948	0.980
	3	0.953	0.947	0.947	0.951	0.972
	4	1.000 (982, 18)	0.981 (834, 166)	0.945 (430, 570)	0.953 (67, 933)	0.980 (7, 993)

Table 2.[5]

$\rho = -0.25$	Test	$\mu_{11}^{(A)}$				
$\mu_{12}^{(A)}$		0	1	2	3	4
0	1	0.044	0.129	0.389	0.751	0.949
	2	0.056	0.166	0.436	0.785	0.953
	3	0.041	0.121	0.367	0.740	0.945
1	1	0.061	0.171	0.447	0.802	0.961
	2	0.070	0.238	0.569	0.868	0.971
	3	0.053	0.135	0.383	0.750	0.947
2	1	0.122	0.250	0.560	0.856	0.970
	2	0.106	0.328	0.691	0.926	0.984
	3	0.117	0.197	0.430	0.769	0.952
3	1	0.250	0.397	0.685	0.906	0.979
	2	0.169	0.442	0.793	0.958	0.990
	3	0.233	0.308	0.521	0.815	0.962
4	1	0.400	0.570	0.793	0.947	0.989
	2	0.245	0.563	0.868	0.972	0.997
	3	0.390	0.450	0.629	0.857	0.966

[5] One table for each ρ considered.

Table 2. continues:

$\rho = 0.0$	Test	$\mu_{11}^{(A)}$				
$\mu_{12}^{(A)}$		0	1	2	3	4
0	1	0.044	0.122	0.377	0.723	0.936
	2	0.051	0.154	0.400	0.739	0.928
	3	0.047	0.128	0.376	0.748	0.948
	4	0.035 (554, 446)	0.121 (684.316)	0.369 (886, 114)	0.715 (966, 34)	0.934 (995, 5)
1	1	0.060	0.142	0.386	0.731	0.937
	2	0.064	0.203	0.487	0.808	0.955
	3	0.056	0.137	0.381	0.746	0.945
	4	0.045 (526, 474)	0.141 (666, 334)	0.391 (859, 141)	0.733 (967, 33)	0.937 (997, 3)
2	1	0.115	0.203	0.446	0.773	0.946
	2	0.080	0.260	0.571	0.868	0.967
	3	0.116	0.192	0.418	0.758	0.947
	4	0.066 (451.549)	0.197 (558.442)	0.445 (788, 212)	0.777 (944, 56)	0.946 (992, 8)
3	1	0.236	0.306	0.549	0.824	0.960
	2	0.116	0.326	0.666	0.912	0.977
	3	0.229	0.294	0.494	0.788	0.951
	4	0.111 (341, 659)	0.264 (447, 553)	0.551 (714, 286)	0.836 (893, 107)	0.962 (973, 27)
4	1	0.381	0.455	0.665	0.878	0.967
	2	0.162	0.409	0.741	0.938	0.986
	3	0.393	0.441	0.605	0.838	0.962
	4	0.161 (237, 763)	0.366 (332, 668)	0.657 (594, 406)	0.879 (819, 181)	0.969 (945, 55)

Table 2. continues:

$\rho = 0.25$	Test	$\mu_{11}^{(A)}$				
$\mu_{12}^{(A)}$		0	1	2	3	4
0	1	0.044	0.140	0.399	0.756	0.957
	2	0.054	0.147	0.418	0.760	0.928
	3	0.045	0.128	0.378	0.750	0.950
	4	0.035 $(536, 464)$	0.130 $(680, 320)$	0.399 $(878, 122)$	0.753 $(974, 26)$	0.957 $(996, 4)$
1	1	0.061	0.129	0.377	0.728	0.942
	2	0.053	0.180	0.464	0.793	0.948
	3	0.061	0.138	0.378	0.744	0.947
	4	0.044 $(534, 466)$	0.136 $(611, 389)$	0.386 $(800, 200)$	0.735 $(954, 46)$	0.943 $(990, 10)$
2	1	0.122	0.171	0.389	0.727	0.936
	2	0.061	0.203	0.517	0.828	0.958
	3	0.117	0.186	0.406	0.748	0.945
	4	0.067 $(462, 538)$	0.155 $(475, 525)$	0.405 $(711, 289)$	0.741 $(886, 114)$	0.937 $(974, 26)$
3	1	0.250	0.250	0.447	0.751	0.937
	2	0.077	0.242	0.566	0.854	0.968
	3	0.219	0.270	0.458	0.765	0.947
	4	0.115 $(366, 634)$	0.198 $(335, 665)$	0.465 $(545, 455)$	0.774 $(787, 213)$	0.942 $(924, 76)$
4	1	0.400	0.397	0.560	0.802	0.949
	2	0.094	0.297	0.608	0.883	0.976
	3	0.383	0.414	0.562	0.805	0.951
	4	0.168 $(293, 707)$	0.278 $(218, 782)$	0.552 $(386, 614)$	0.815 $(646, 354)$	0.952 $(841, 159)$

Table 2. continues:

$\rho = 0.50$	Test	$\mu_{11}^{(A)}$				
$\mu_{12}^{(A)}$		0	1	2	3	4
0	1	0.044	0.161	0.481	0.865	0.992
	2	0.055	0.166	0.501	0.828	0.969
	3	0.046	0.128	0.379	0.752	0.954
	4	0.048 (498, 502)	0.155 (656, 344)	0.486 (889, 111)	0.864 (979, 21)	0.992 (998, 2)
1	1	0.065	0.122	0.407	0.793	0.983
	2	0.053	0.162	0.492	0.831	0.971
	3	0.061	0.135	0.374	0.747	0.950
	4	0.049 (492, 508)	0.126 (526, 474)	0.431 (784, 216)	0.800 (932, 68)	0.983 (987.13)
2	1	0.149	0.151	0.377	0.742	0.961
	2	0.050	0.166	0.497	0.839	0.972
	3	0.123	0.180	0.393	0.741	0.946
	4	0.083 (471, 529)	0.138 (390, 610)	0.410 (602, 398)	0.769 (833, 167)	0.963 (947, 53)
3	1	0.294	0.232	0.397	0.723	0.942
	2	0.053	0.169	0.494	0.837	0.973
	3	0.219	0.258	0.439	0.753	0.945
	4	0.142 (443, 557)	0.154 (258.742)	0.427 (404, 596)	0.765 (662.338)	0.947 (858.142)
4	1	0.489	0.371	0.474	0.733	0.936
	2	0.057	0.173	0.491	0.828	0.971
	3	0.374	0.391	0.518	0.775	0.947
	4	0.212 (408, 592)	0.184 (173, 827)	0.454 (233, 767)	0.770 (472, 528)	0.951 (707, 293)

Table 2. continues:

$\rho = 0.75$	Test	$\mu_{11}^{(A)}$				
$\mu_{12}^{(A)}$		0	1	2	3	4
0	1	0.044	0.244	0.738	0.989	1.000
	2	0.045	0.263	0.734	0.977	1.000
	3	0.044	0.127	0.378	0.753	0.956
	4	0.033 (478, 522)	0.244 (725, 275)	0.736 (951, 49)	0.990 (991, 9)	1.000 (998, 2)
1	1	0.088	0.144	0.559	0.952	1.000
	2	0.056	0.188	0.643	0.958	1.000
	3	0.054	0.125	0.369	0.745	0.951
	4	0.063 (527, 473)	0.146 (523.477)	0.578 (839, 161)	0.954 (966, 34)	1.000 (993, 7)
2	1	0.238	0.135	0.425	0.880	0.999
	2	0.087	0.128	0.548	0.926	0.998
	3	0.113	0.154	0.370	0.740	0.946
	4	0.163 (599, 401)	0.111 (344, 656)	0.459 (636, 364)	0.892 (876.124)	0.999 (969, 31)
3	1	0.477	0.229	0.377	0.793	0.991
	2	0.145	0.089	0.438	0.883	0.996
	3	0.219	0.239	0.402	0.740	0.945
	4	0.334 (644, 356)	0.102 (274, 726)	0.377 (358, 642)	0.826 (688.312)	0.991 (898.102)
4	1	0.728	0.429	0.413	0.729	0.979
	2	0.205	0.068	0.334	0.819	0.991
	3	0.365	0.370	0.473	0.749	0.945
	4	0.523 (687, 313)	0.152 (268, 732)	0.309 (163, 837)	0.768 (407, 593)	0.982 (740, 260)

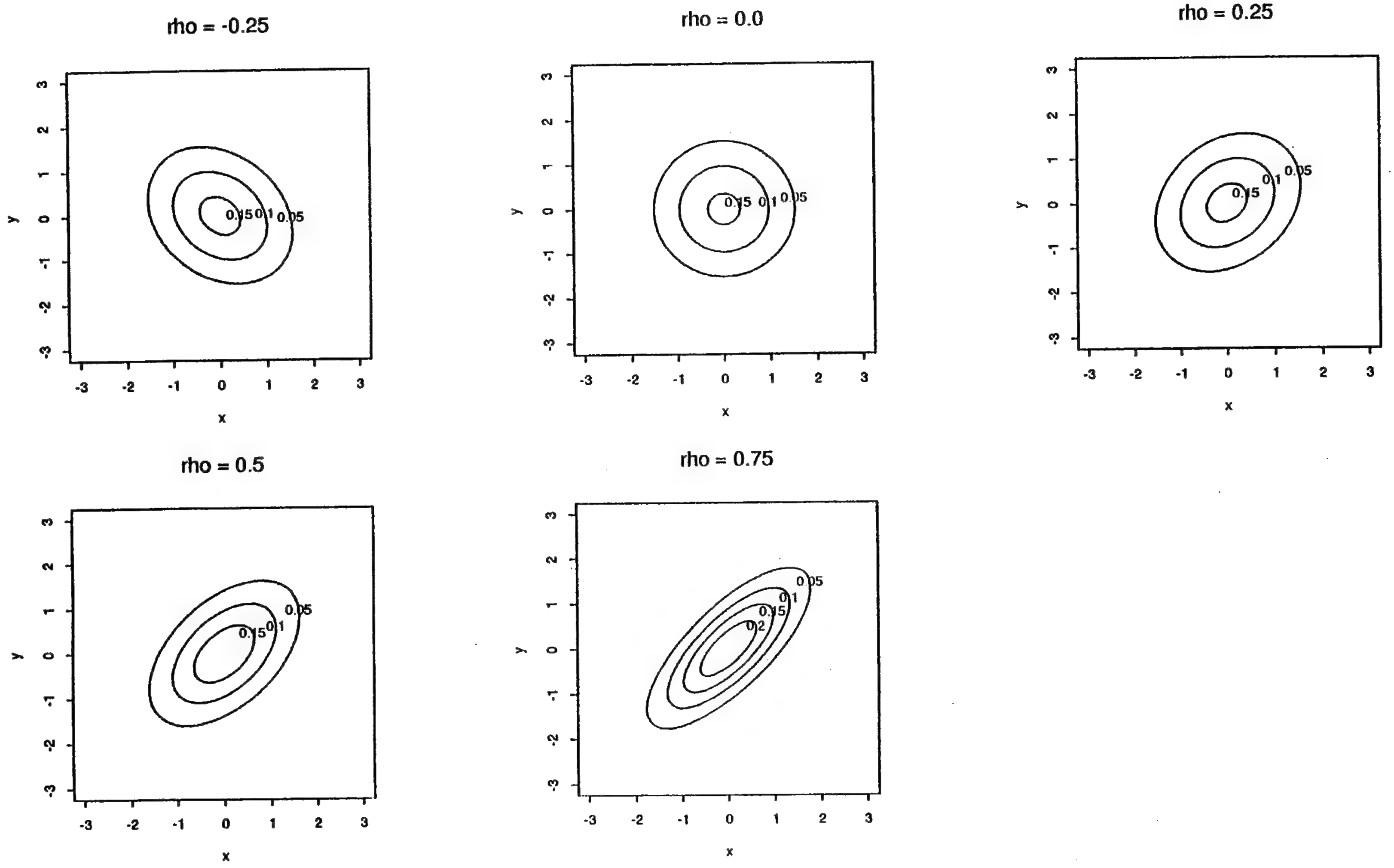


Figure 1. Contour plot of Bivariate Normal Density Function:
mu=(0,0); s1=1, s2=1, rho=-0.25,0.0,0.25,0.5,0.75.

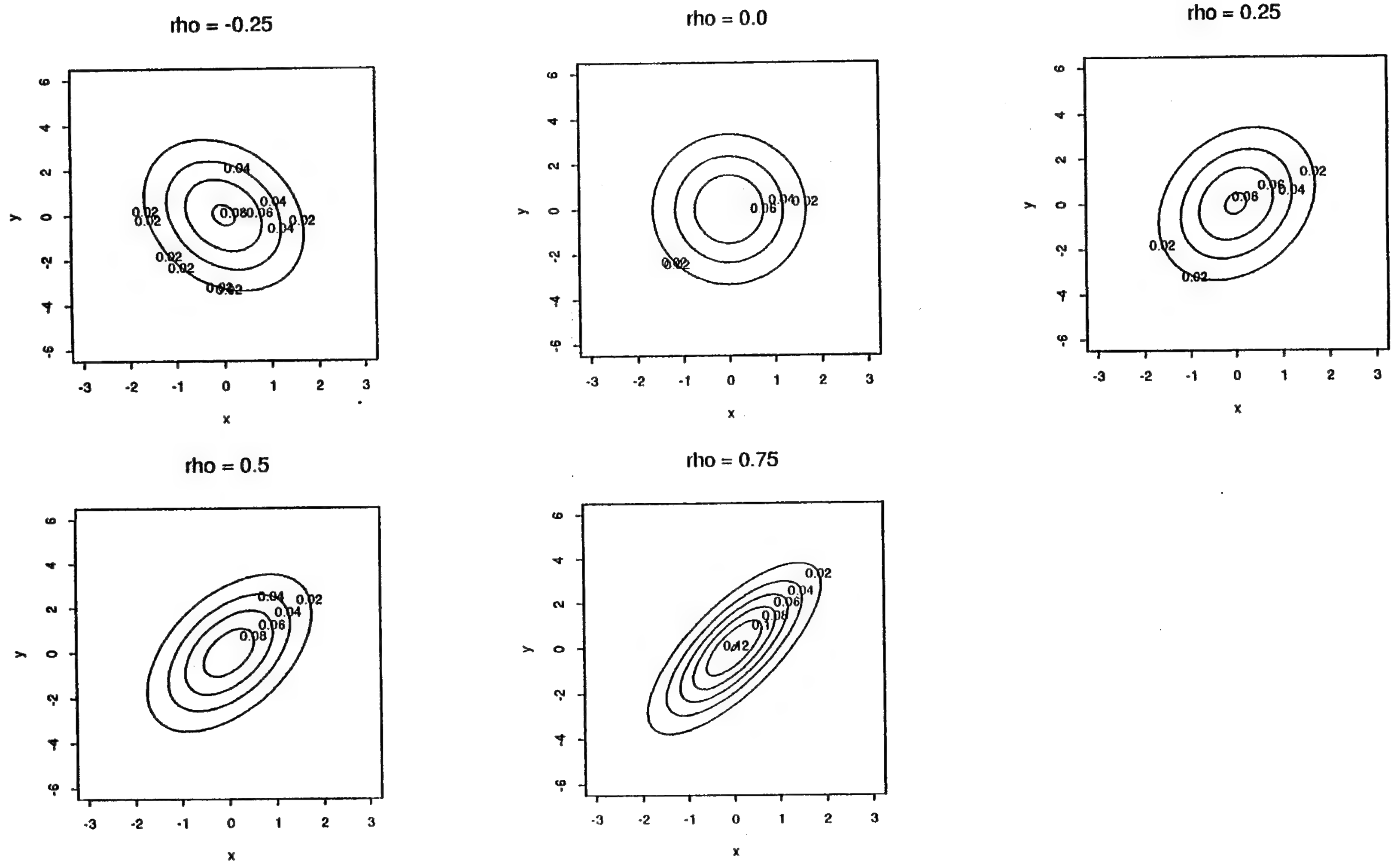


Figure 2. Contour plot of Bivariate Normal Density Function: mu=(0,0); s1=1, s2=4, rho=-0.25,0.0,0.25,0.5,0.75.

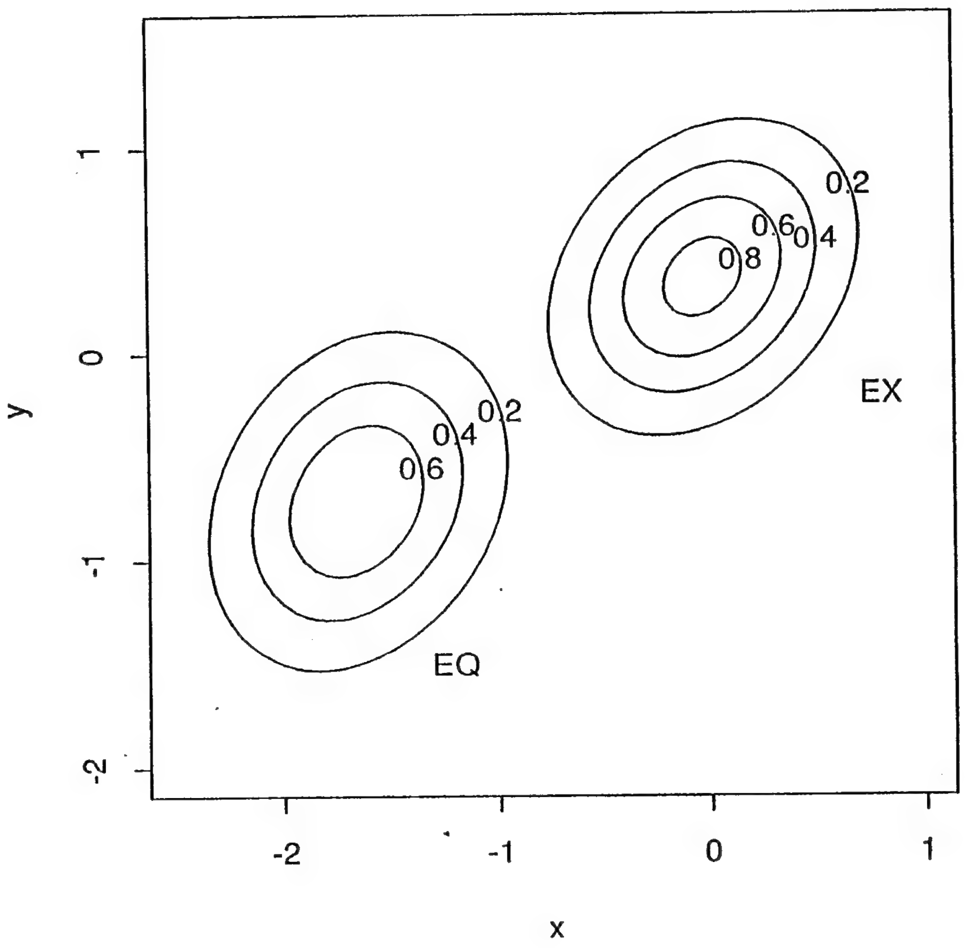


Figure 3: Contours for explosions and earthquakes using log(Pn/Lg ratio -- 6-8 Hz) at KNB and MNV

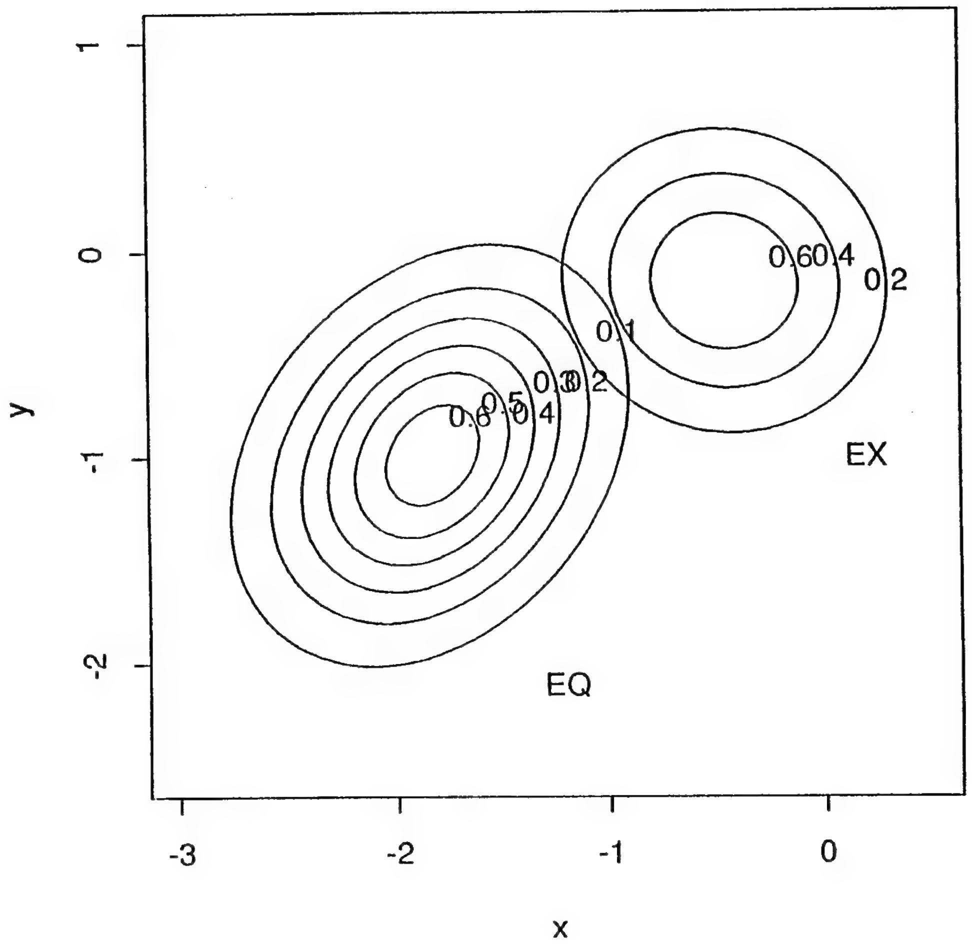


Figure 4: Contours for explosions and earthquakes using log(Pn/Lg -- 4-6 Hz) at KNB and MNV

4 Concluding Remarks

We considered the problem of observing seismic events for the purpose of distinguishing between earthquakes and explosions. We studied the case in which data are available at more than one station. We discussed three outlier tests, all based on likelihood ratio, for the cases in which there are p feature variables at each of m stations. The tests utilize the data in different forms: Test 1 treats p variables at m stations as mp variables, Test 2 combines the information for each of p variables at m stations, Test 3 treats each station separately. We also discussed a combined test, Test 4, which decides to use Test 1 or Test 2 based on Mahalanobis distance criterion described in section 1.

It seems that Test 1 gives powers that are sometimes best and always comparable to the best powers, while for the other tests, scenarios existed in which the power was substantially lower than that for Test 1. Thus, unless m and p are sufficiently large to make Test 1, the full vector test, impractical, we recommend its use.

5 Acknowledgment

The research was partially supported by ARPA Contract No. F19628-93-C0199.

A New Test for Outlier Detection from a Multivariate Mixture Distribution

Suojin Wang, Wayne A. Woodward, H.L. Gray, and Stephen Wiechecki

ABSTRACT

The problem of testing an outlier from a multivariate mixture distribution of several populations has many important applications in practice. One particular example is in monitoring worldwide nuclear testing, where we wish to detect whether an observed event is possibly a nuclear explosion (an outlier) by comparing it with the training samples from mining blasts and earthquakes. The combined population of seismic events from mining blasts and earthquakes can be viewed as a mixture of two populations. The classical likelihood ratio test appears to be not applicable in our problem, and in spite of the importance of this problem, little progress has been made in the literature. In this report we propose a simple modified likelihood ratio test that overcomes the difficulties in the current problem. Bootstrap techniques are used to approximate the distribution of the test statistic. The advantages of the new test are demonstrated via simulation studies. Some new computational findings are also reported.

This research was partially supported by ARPA Contracts F19628-93-C-0199 and F19628-95-C-0098, NSF Grant DMS-9504589, an NSA grant and Texas A&M Scholarly and Creative Activities Program 95-59. We would like to acknowledge Steve Sain for producing the graphics for the figures.

1. Introduction

An extremely important practical problem is that of monitoring worldwide nuclear testing, where we wish to detect whether an observed seismic event may be a nuclear explosion by comparing it with the training samples obtained from previous seismic activity in the region. In this case, the training data will often be composed of data which are a composite of mining explosions and earthquakes. Usual methods of outlier detection typically focus on the setting in which observations are tested as outliers from a single population. However, in the case considered here, there are two populations, and we wish to test whether a seismic event should be considered to be an outlier from either or both of the populations. Actually, these results are applicable to two or more populations but we focus on the case of two. Another point of interest is the fact that the setting considered here differs from a common outlier scenario in which a sample is given and the observations from the sample are tested to determine whether they should be considered as outliers from the population from which the sample was obtained. This, however, is not the scenario considered here. Specifically, in our setting, "pure" samples from the populations in question are available, and our desire is to test a new observation as an outlier from these populations. We will refer to this testing procedure as outlier testing throughout the report.

The classical method for outlier detection of the type we are addressing is the likelihood ratio test (Wilks (1963), Caroni and Prescott (1992)), usually under the normality assumption for the multivariate distributions of the training sample population and the outlier population, and under the assumption of equal covariance of the two populations under the alternative hypothesis. The resulting test is essentially the Hotelling's T^2 test (see Anderson (1984)). In our current problem, because of the fact that there is not a single multivariate normal population associated with the training sample, these assumptions are not satisfied. Thus, a direct application of the standard likelihood ratio test does not seem possible. In spite of the importance of this problem, to our knowledge little progress has been made in the literature. Baek et al. (1992) recently considered the outlier testing in the seismic setting discussed here but in the special case in which seismic events are tested as outliers from a single population, usually earthquakes. Baek et al.

(1992) used a bootstrap approach to ascertain the distribution of the likelihood ratio when the multivariate distribution associated with the training sample has both continuous components and discrete components that have a finite number of possible outcomes. Some assumptions, such as equality of covariances, are imposed to link the training sample population and the outlier population. It is possible to apply the test of Baek et al. sequentially to each training sample population, but this can be cumbersome, e.g. the training sample populations often have different covariance structures. Furthermore, this procedure would result in substantial loss of power.

In this report we consider an approach to the practical problem at hand by considering the combined population of seismic events of mining blasts and earthquakes as a mixture of two populations. We propose a simple modified likelihood ratio test using bootstrap resampling that appears to perform well in this setting. The methodology is presented in Section 2 for testing outliers from a mixture population consisting of m components. Some numerical procedures are addressed, including the use of the bootstrap for approximating the distribution of the test statistic in Section 3. We also describe how the intensive computing time required for the bootstrap resampling can be reduced without loss of accuracy when the training sample size is relatively large. Section 4 provides the results of empirical studies. Some concluding remarks are given in Section 5.

2. The Methodology

Suppose we have a mixture distribution Π of m populations, Π_i, $i = 1, \ldots, m$. In the nuclear testing example mentioned above, $m = 2$ for mining explosions and earthquakes. Let d be the dimension of the variables from the mixed population Π, and for clarity in the presentation assume all the distributions are continuous. Note that extensions to discrete or mixed cases are mainly a matter of notational adjustments. The density of the mixture distribution is

$$f(\boldsymbol{x};\, \boldsymbol{\theta}) = \sum_{i=1}^{m} p_i g_i\,(\boldsymbol{x};\, \theta_i)\,, \tag{1}$$

where $p_i \geq 0$ are mixing proportions with $\sum_{i=1}^{m} p_i = 1$, g_i are the densities of Π_i, θ_i are unknown parameter vectors, $\boldsymbol{\theta} = (p_1, ..., p_m, \theta_1', ..., \theta_m')'$ and $\boldsymbol{x} = (x_1, ..., x_d)'$. In the nuclear monitoring scenario, we wish to test whether a new seismic event is an outlier to the mixture of earthquakes and mining explosions. More generally, we wish to be able to test whether a new observation is an outlier from the mixed population Π.

Assume that we have a random training sample of size n from the mixture population

$$\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \Pi,$$

and that we are able to identify the associated source population for $n_L \leq n$ members of the training sample. For convenience, let

$$\boldsymbol{X}_{k_{i-1}+1}, \boldsymbol{X}_{k_{i-1}+2}, ..., \boldsymbol{X}_{k_i} \in \Pi_i, \text{ for } i = 1, ..., m, \tag{2}$$

where $0 = k_0 < k_1 < ... < k_m = n_L$, i.e., $n_i = k_i - k_{i-1}$ (normally ≥ 10) data points are identified to be from Π_i. Additionally, we allow for the possibility that the training sample contains n_U unlabeled observations from the mixture. In the notation of Redner and Walker (1984) we assume the sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ is of Type 4, i.e. the training sample consists of labeled and unlabeled observations. The associated n_i's, $i = 1, ..., m$ are random variables following a multinomial distribution, and they contain information about the mixing proportions. In this notation, $n = n_L + n_U$. If in fact $n_U = 0$, then the training sample consists of only labeled observations and is a sample of Type 3 using the Redner and Walker notation. Now a new observation $\boldsymbol{X}_{n+1}$ is obtained. Given (2) we want to test the following hypotheses:

$$H_0 : \boldsymbol{X}_{n+1} \in \Pi$$

vs. (3)

$$H_1 : \boldsymbol{X}_{n+1} \notin \Pi .$$

The classical likelihood ratio test statistic is the ratio of the maximized likelihood functions under H_0 and H_1. Under H_0 the sample is of Redner and Walker Type 4, i.e. we assume that

$\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are as before while $\boldsymbol{X}_{n+1}$ is unlabeled but from the same mixture distribution as $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$. That is, we assume that all $n+1$ observations are from the mixture distribution assumed under H_0 with n_L of these labeled and n_U+1 unlabeled. The likelihood function under H_0 is

$$L_0(\theta) = \frac{n_L!}{n_1! \ldots n_m!}\left(\prod_{i=1}^{m} \prod_{j=k_{i-1}+1}^{k_i} p_i g_i(\boldsymbol{X}_j; \theta_i)\right)\left(\prod_{s=n_L+1}^{n} f(\boldsymbol{X}_s; \boldsymbol{\theta})\right) f(\boldsymbol{X}_{n+1}; \boldsymbol{\theta}).$$

Let $h(\boldsymbol{x};\, \alpha)$ be the density associated with the outlier population from which $\boldsymbol{X}_{n+1}$ is sampled, where α is an unknown parameter vector. Then the likelihood function under H_1 is

$$L_1(\theta,\alpha) = \frac{n_L!}{n_1! \ldots n_m!}\left(\prod_{i=1}^{m} \prod_{j=k_{i-1}+1}^{k_i} p_i g_i(\boldsymbol{X}_j; \theta_i)\right)\left(\prod_{s=n_L+1}^{n} f(\boldsymbol{X}_s; \boldsymbol{\theta})\right) h(\boldsymbol{X}_{n+1};\, \alpha). \qquad (4)$$

Difficulties arise when maximizing L_1 since there is only a single observation from the outlier population so that generally no suitable MLE is possible for α, unless α is assumed to directly link to θ. Any such linkage assumption is quite questionable since we now have m individual populations that make up the mixture distribution. Furthermore, with only one observation it is impossible to do any model checking of $h(\boldsymbol{x};\, \alpha)$. To overcome these difficulties and to observe the fact that little information is known about the outlier population from which $\boldsymbol{X}_j$ is sampled, we simply use a constant density $h(\boldsymbol{x}) \equiv c$ over its practical (finite) support. Moreover, the constant density is also assumed in the bootstrap procedure described below. Thus, dropping the constant from the likelihood ratio test statistic will not affect any test conclusions. Therefore we let

$$\widetilde{L}_1(\theta) = \frac{n_L!}{n_1! \ldots n_m!}\left(\prod_{i=1}^{m} \prod_{j=k_{i-1}+1}^{k_i} p_i g_i(\boldsymbol{X}_j; \theta_i)\right)\left(\prod_{s=n_L+1}^{n} f(\boldsymbol{X}_s; \boldsymbol{\theta})\right),$$

which is the likelihood based on the sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ from the mixture. We define a simple modified likelihood ratio test statistic

$$W = \frac{\sup_{\theta \in \Theta} L_0(\theta)}{\sup_{\theta \in \Theta} \widetilde{L}_1(\theta)}, \tag{5}$$

where Θ is the entire parameter space. It is easily seen that the departure of $\boldsymbol{X}_{n+1}$ from f will reduce $\sup_{\theta \in \Theta} L_0(\theta)$ making W small. Hence the rejection region is of the form $W \leq W_\alpha$ for some W_α picked to provide a level α test. Since the null distribution of W has no known closed form, we suggest the use of the parametric bootstrap method to approximate it, as shown in the next section. Based on the discussion here the use of W seems to be a reasonable approach, and in Section 4 we demonstrate that W performs well under all the simulation scenarios considered.

Concluding this section, we point out that asymptotically $W \approx f(\boldsymbol{X}_{n+1}; \widehat{\theta}_n)$, as $n \rightarrow \infty$, where $\widehat{\theta}_n$ is the MLE using the training sample only. See the Appendix for the proof. Moreover, the bootstrap-one method described in the next section is essentially equivalent to using this asymptotic result.

3. The Bootstrap and Other Computational Procedures

In this section we discuss numerical issues associated with the test procedure described in Section 3. It should be noted that often both the numerator and denominator of W in (5) may be difficult to obtain since the individual densities are mixture distributions. Recall also that for the numerator we assume that $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ can be identified with their component population, but $\boldsymbol{X}_{n+1}$ is only known to be from the mixture, not the exact component. However, if we consider the setting of multivariate normality for each component, i.e.,

$$g_i(\boldsymbol{x};\, \theta_i) \sim N(u_i, \Sigma_i)\,, \tag{6}$$

and thus $f(\boldsymbol{x};\, \theta)$ is a mixture of m multivariate normal distributions, a numerical iteration algorithm based on the EM algorithm has been developed by Redner and Walker (1984), for maximizing $L_0(\theta)$. They extended Hosmer's (1973) algorithm for the case of two univariate normal components to the multivariate normal components setting, and in our simulation studies, we have adapted their method. Note that with (6), $\sup_{\theta \in \Theta} \widetilde{L}_1(\theta)$ is easily obtained. Using the resulting estimator $\widehat{\theta}_n$ as an initial value in the numerator, it only takes at most a few steps to obtain convergence.

We now turn to bootstrapping the null distribution of W. We will employ the parametric bootstrap based on the training sample $X_1, ..., X_n$. The following algorithm is used which mimics the original sampling plan.

Step 1: Use (2) to obtain $(\widehat{p}_i, \widehat{u}_i, \widehat{\Sigma}_i)$ for $i = 1, ..., m$.

Step 2: For each integer b, $b = 1, ..., B$, draw a sample of size n_L from the multinomial distribution with $\widehat{\boldsymbol{p}} = (\widehat{p}_1, \dots, \widehat{p}_m)'$. We observe the frequencies n^b_{1L}, n^b_{2L}, ..., and n^b_{mL} where $n^b_{1L} + n^b_{2L} + \cdots + n^b_{mL} = n_L$. Additionally, we draw a sample of size n_U from the same multinomial distribution resulting in frequencies n^b_{1U}, n^b_{2U}, ..., and n^b_{mU} where $n^b_{1U} + n^b_{2U} + \cdots + n^b_{mU} = n_U$.

Step 3: Draw samples of size $n^b_{iL}$and n^b_{iU} from $N(\widehat{u}_i, \widehat{\Sigma}_i)$ for $i = 1, ..., m$. The n_L observations associated with frequencies n^b_{iL}, $i = 1$, ..., m are treated as labeled samples in the analysis, while the n_U observations corresponding to n^b_{iU}, $i = 1$, ..., m are treated as unlabeled observations. These resampled data are used to compute the test statistic in (5). This test statistic is denoted by W^*_b.

Step 4: Draw a new, $(n+1)$st, observation from the empirical mixture by randomly selecting a single observation from the multinomial distribution in Step 2. This multinomial will essentially select a component i between 1 and m, and we generate an observation from the associated $N(\widehat{u}_i, \widehat{\Sigma}_i)$ distribution.

Step 5: Repeat Steps 2 to 4 B times $(b = 1, ..., B)$. Then define W_α to be the (100α)th percentile of all W^*_b. Specifically, if $\alpha = j/(B+1)$, then W_α is the jth smallest value of $\{W^*_b\}^B_{b=1}$ (see McLachlan, 1987). Statistical decisions can then be made.

Notice that when n is large the bootstrap scheme may require considerable computing time. However, when n_i are not very small, this computational burden can be avoided by employing an approximate bootstrap scheme, called bootstrap-one. This technique uses the original training sample in Steps 2 and 3 for all $b = 1, ..., B$. It effectively eliminates these two steps and many calculations in obtaining W_α.

The bootstrap-one method conceptually approximates the conditional distribution of W given $X_1, ..., X_n$. When all n_i are relatively large, the conditioning effect is minimal. The accuracy and advantages of the bootstrap-one method are among the things studied in simulations which are discussed in the next section.

4. Empirical Studies

In this section we report some results of a simulation study to illustrate the performances of the new methods. In these simulations we focus on the case in which all training sample observations are labeled, i.e. $n_U = 0$.

Example 1. In this example, we choose $m = 1$, $d = 2$, and $n = 40$ so that the training sample is from a bivariate $N(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix} \tag{7}$$

were used. Obviously, in this case since there is only one component in the "mixture", all observations in the training sample can be labeled, i.e. $n_U = 0$. The reason for choosing $m = 1$ is that in this case it is easy to apply the standard likelihood ratio test assuming that the outlier population is normal with the same covariance Σ. In this case, there is a single training sample of size n and an observation X_{n+1} to be tested as an outlier. Baek et al. (1992) discusses the generalized likelihood ratio test in this setting. In particular, the likelihood ratio statistic is given by

$$\lambda = \frac{\sup\limits_{\theta \in \Theta} L_0(\theta)}{\sup\limits_{\theta \in \Theta} L_1(\theta, \alpha)}\ , \tag{8}$$

where $L_1(\theta,\ \alpha)$ is given in (4) and α is related to θ in a certain way.

Specifically, h is the multivariate normal density associated with observation $\boldsymbol{X}_{n+1}$ and $\alpha = (\mu_2, \Sigma)$ is estimated by taking $\hat{\mu}_2 = \boldsymbol{X}_{n+1}$and taking $\widehat{\Sigma}$ to be the MLE obtained from the training sample. Under the normality assumption in this example, the test statistic in (8) is known to be distributed as Hotelling's T^2 (e.g. Anderson, 1984). Baek et al. (1992) considered the likelihood ratio in (8), where the multivariate random variables could be composed of both continuous and discrete components. They approximated the distribution of λ in this case using the bootstrap procedure described here. They applied the bootstrap procedure to the special case in which the distributions were multivariate normal and approximated the distribution of λ using the bootstrap procedure. Simulations have shown that the power of the test based on the bootstrap is very similar to that obtained based on Hotelling's T^2 in the multivariate normal case. In this report all tests are based on the use of bootstrap resampling to approximate the distribution of the test statistic. The test based on (8) will be called the "standard" likelihood ratio test.

Instead of including $L_1(\theta,\ \alpha)$ in the denominator of (8) in this multivariate normal setting, we could have used the test statistic given in (5) which is based on the use of a constant density $h(x) = c$ for over its support. The test statistic using (5) will be termed the "modified" likelihood ratio test. For each of these tests, whenever we approximate the distribution of the test statistic by a full bootstrapping of $n+1$ observations, we will refer to this as the "full" procedure. Alternatively, in each case we also consider the use of the bootstrap-one technique. In Table 1 we denote them as "full" and "one" respectively.

Table 1 summarizes the simulation results of the two tests. One thousand replications were used for each entry and we used $B = 499$. The power was obtained with $N\left(\binom{2}{2},\right.$

$\begin{pmatrix} 1 & -.5 \\ -.5 & 1 \end{pmatrix}\bigg)\bigg)$ as the outlier population. We have experimented with other covariance values, including that in (7), and similar power patterns were observed.

First, we compare the standard and modified tests using full bootstrapping. In Table 1, it can be seen that the significance levels for both tests are close to the nominal level of $\alpha = .05$ with the modified tests having slightly larger levels. Additionally, the powers of the two tests are similar with the modified tests having somewhat larger power. Thus, the use of W in (5), which appropriately reflects our ignorance about the outlier population, performs as well as the full likelihood ratio.

Next, comparing "One" columns to "Full" columns, we observe that the bootstrap-one has significance levels that are artificially high for smaller sample sizes. However, for large n (say $\geq$ 100) the significance levels are of appropriate size. For these larger sample sizes the bootstrap-one procedure tended to have higher power than obtained using full bootstrapping. Based on these results and the computational burden associated with large n suggests that the bootstrap-one is a viable alternative. Finally, notice that the bootstrap-one method is identical for the standard and new tests. In fact, the identity can be shown analytically under normality. However, the identity is not true in general.

Example 2. In this example we consider the use of the likelihood ratio test to test for outliers from the mixture model in (1) with $m = 2$ and $n = 60$. Again we consider the case in which $d = 2$ and $n_U = 0$, and specifically, we assume that the component densities g_1 and g_2 are multivariate normal densities associated with a

$$N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}\right)$$

and

$$N\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & -.5 \\ -.5 & 1 \end{pmatrix}\right)$$

populations respectively.

Case a: $p_1 = p_2 = .5$.

We examine the power of the test for detecting outliers from

$$N\left(\begin{pmatrix}1+k-5\\1-(k-5)\end{pmatrix}, \begin{pmatrix}1 & .5\\.5 & 1\end{pmatrix}\right)$$

population where $k = 1, ..., 9$. In Figure 1(a) we show data from a mixture of two populations with $p_1 = p_2 = 0.5$ along with 5 outliers. In Figure 1(b) we show the same data with individual observations labeled with regard to the associated component population or outlier population. The outliers are indicated by solid dots. In Figure 1(c) we again show the labeled data along with contours of the mixture population. Finally, in Figure 1(d) we show means and contours of the two component populations and of the outlier population. In Figure 2 we show the contours of the mixture components as in Figure 1(d) along with the outlier means $(1+k-5, 1-(k-5))'$, $k = 1, ..., 9$. Also in this figure we show the contour of the outlier population for the case $k = 2$, i.e. the mean is $(-2, 4)'$. In Table 2(a) $n = 60$ is used and the nominal level is $\alpha = 0.05$. As can be seen, the significance level is close to the nominal level. Whenever the outlier population is well separated from the component distributions of the mixture we have good power while as would be expected the power lowers dramatically for k near 5. The true powers for $k = 1, 2, 3,$ and 4 are the same as those for $k = 9, 8, 7$ and 6 respectively, due to symmetry. The empirical results appear to verify this fact.

Case b: $p_1 = 0.25$ and $p_2 = 0.75$.

In this case we consider the same scenario as Case a but with $p_1 = 0.25$ and $p_2 = 0.75$. In Figure 3 we show the plots corresponding to Figure 1 for the case in which $p_1 = 0.25$ and $p_2 = 0.75$, and in Table 2(b) we show results corresponding to those in Table 2(a) for this case. Again, we see that the significance levels are accurate and that powers are similar to those in Table 2(a). It should be noted that due to smaller p_1 here, there was a very small fraction

(< 0.2%) of all bootstrap simulation replications that did not converge with our current program. This problem seems to become more serious when smaller values of n are used. In our analysis we simply skip any bootstrap realization for which convergence was not obtained and generate another one. Another possible approach would be to use the starting values as final estimates for these bootstrap replications.

5. Concluding Remarks

In this report we have proposed a simple modified likelihood ratio test for multivariate outlier detections. This new test is not only good for use in general outlier detection problems, but especially applicable when the training sample population is a mixture of several populations. In the new test no assumption is necessary for the covariance structure or any other moments of the outlier population, and in fact no parametric modeling is required for the outlier population. Furthermore, although with weaker assumptions it is more powerful than the standard likelihood ratio test in the simpler non mixture situation in which the standard test applies.

We have also investigated bootstrapping the distributions of the test statistics. The computationally intensive resampling method seems to be quite effective. When the training sample size is large, we have also suggested the bootstrap-one method, which significantly reduces the computing time and seems to have somewhat more power.

It should be noted that the procedure could be extended to cover the case in which all of the training sample observations are unlabeled. This, however, will require dealing with issues such as the use of appropriate starting values and is not considered here.

APPENDIX

In this appendix, we show that $W \approx f(\boldsymbol{X}_{n+1};\, \widehat{\theta}_n)$, as $n \to \infty$, where W is given in (5). Let

$$\widetilde{\ell}_1(\theta) = \ln\{\widetilde{L}_1(\theta)\},$$

$$\ell_0(\theta) = \ln\{L_0(\theta)\} = \widetilde{\ell}_1(\theta) + \ln\{f(\boldsymbol{X}_{n+1}; \theta)\}. \qquad (A1)$$

Suppose $\widehat{\theta}_n$ and $\widehat{\theta}_{n+1}$ satisfy the conditions that $\widetilde{L}_1(\widehat{\theta}_n) = \sup_{\theta \in \Theta} \widetilde{L}_1(\theta)$ and $L_0(\widehat{\theta}_{n+1}) = \sup_{\theta \in \Theta} L_0(\theta)$, respectively. Then $\ell_0'(\widehat{\theta}_{n+1}) = 0$ and $\widetilde{\ell}_1'(\widehat{\theta}_n) = 0$. Thus, from

$$\begin{aligned}\ell_0'(\widehat{\theta}_{n+1}) &= \ell_0'(\widehat{\theta}_n) + \ell_0''(\widehat{\theta}_n)(\widehat{\theta}_{n+1} - \widehat{\theta}_n) + \text{smaller terms} \\ &= \tfrac{\partial}{\partial\theta}[\ln\{f(\boldsymbol{X}_{n+1};\, \theta)\}]\Big|_{\theta=\widehat{\theta}_n} + \ell_0''(\widehat{\theta}_n)(\widehat{\theta}_{n+1} - \widehat{\theta}_n) + \text{smaller terms,}\end{aligned}$$

we have

$$\widehat{\theta}_{n+1} - \widehat{\theta}_n = O_p(\tfrac{1}{n}), \qquad (A2)$$

since $\ell_0'(\widehat{\theta}_{n+1}) = 0$, $\ell_0''(\widehat{\theta}_n)$ is of order $O_p(n)$, and $\frac{\partial}{\partial\theta}[\ln\{f(\boldsymbol{X}_{n+1};\, \theta)\}]\Big|_{\theta=\widehat{\theta}_n}$ is $O_p(1)$. Now by (A1) and (A2),

$$\begin{aligned}W &= \exp\{\ell_0(\widehat{\theta}_{n+1}) - \widetilde{\ell}_1(\widehat{\theta}_n)\} \\ &= \exp\{\ell_0(\widehat{\theta}_n) + \ell_0'(\widehat{\theta}_n)(\widehat{\theta}_{n+1} - \widehat{\theta}_n) + O_p(\tfrac{1}{n}) - \widetilde{\ell}_1(\widehat{\theta}_n)\} \\ &= \exp[\ln\{f(\boldsymbol{X}_{n+1};\, \widehat{\theta}_n)\}] + O_p(\tfrac{1}{n}) \\ &= f(\boldsymbol{X}_{n+1};\, \widehat{\theta}_n) + O_p(\tfrac{1}{n}),\end{aligned}$$

completing the proof.

References

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

Baek, J. Gray, H.L., McCartor, G.D., and Woodward, W.A.(1992). "A Generalized Likelihood Ratio Test in Outlier Detection or Script Matching," Advanced Research Projects Agency Technical Report.

Caroni, C. and Prescott, P. (1992). "Sequential Application of Wilks' Multivariate Outlier Test." *Appl. Statist.*, 41, 355-364.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife." *Ann. Statist.* 7, 1-26.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Hosmer, D.W., Jr. (1973). "A Comparison of Iterative Maximum-Likelihood Estimates of the Parameters of a Mixture of Two Normal Distributions Under Three Different Types of Sample." *Biometrics*, 29, 761-770.

McLachlan, G.J. (1987). "On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture," *Applied Statistics*, 36, 318-324.

Redner, R.A. and Walker, H.F. (1984). "Mixture Densities, Maximum Likelihood and the EM Algorithm." *SIAM Review*, 26, 195-239.

Wilks, S.S. (1963). "Multivariate Statistical Outliers." *Sankhya*, 25, 407-426.

Table 1. Comparisons of significant level and power of the standard likelihood ratio test and modified likelihood ratio test, using two (Full and One) bootstrap approximations.

n	Significance Level				Power			
	Standard		Modified		Standard		Modified	
	Full	One	Full	One	Full	One	Full	One
15	.048	.118	.065	.118	.522	.729	.568	.729
20	.048	.100	.063	.100	.541	.709	.588	.709
25	.036	.081	.048	.081	.563	.704	.601	.704
30	.047	.084	.051	.084	.579	.718	.609	.718
50	.046	.064	.050	.064	.626	.696	.645	.696
100	.056	.059	.057	.059	.646	.677	.657	.677
150	.059	.057	.061	.057	.655	.703	.665	.703
s.e.	.007				.015			

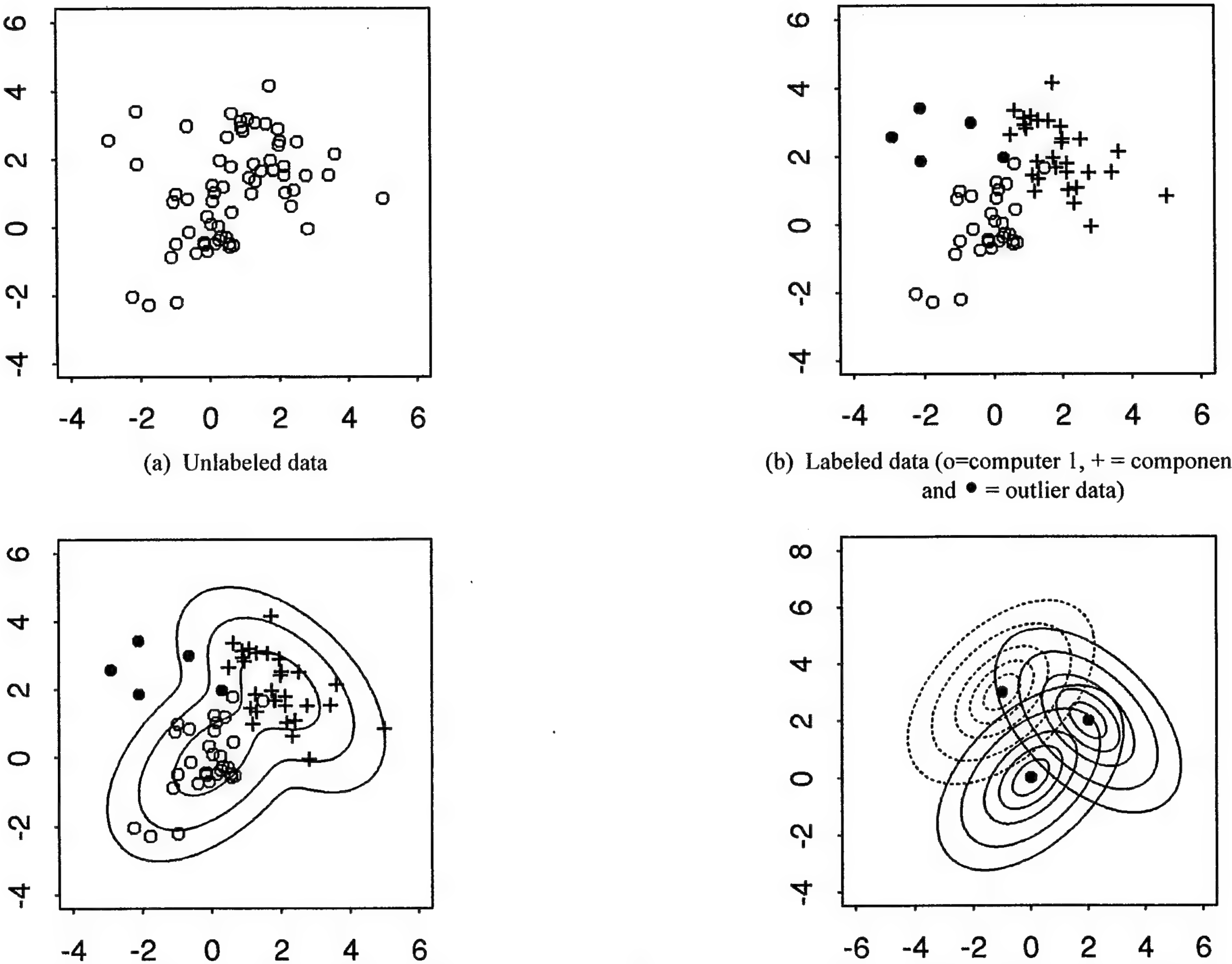
Table 2a. Significance level and power of new test in Example 2; $p_1 = p_2 = 0.5$, $n = 60$, $B = 199$, 1000 replications

Level	.050 (s.e. .007)								
k	1	2	3	4	5	6	7	8	9
Power	1.000	.984	.754	.226	.031	.231	.767	.980	1.000
s.e.	.001	.004	.014	.013	.006	.013	.014	.004	.001

Table 2b. Significance level and power of new test in Example 2; $p_1 = 0.25$, $p_2 = 0.75$, $n = 60$, $B = 199$, 1000 replications

Level	.055 (s.e. .007)								
k	1	2	3	4	5	6	7	8	9
Power	.999	.970	.709	.245	.042	.242	.701	.972	.999
s.e.	.001	.005	.014	.014	.006	.014	.014	.005	.001

Figure 1 - Mixture distributions for Example 2a where $f(x) = .5N(\mu_1, \Sigma_1) + .5N(\mu_2, \Sigma_2)$



(a) Unlabeled data

(b) Labeled data (o=computer 1, + = component 2, and • = outlier data)

(c) Labeled data long with contours of mixture population

(d) Contours of component populations and of outlier population (dotted)

Figure 2 - Mixture distributions for Example 2*a* showing means of outlier distributions in the simulations

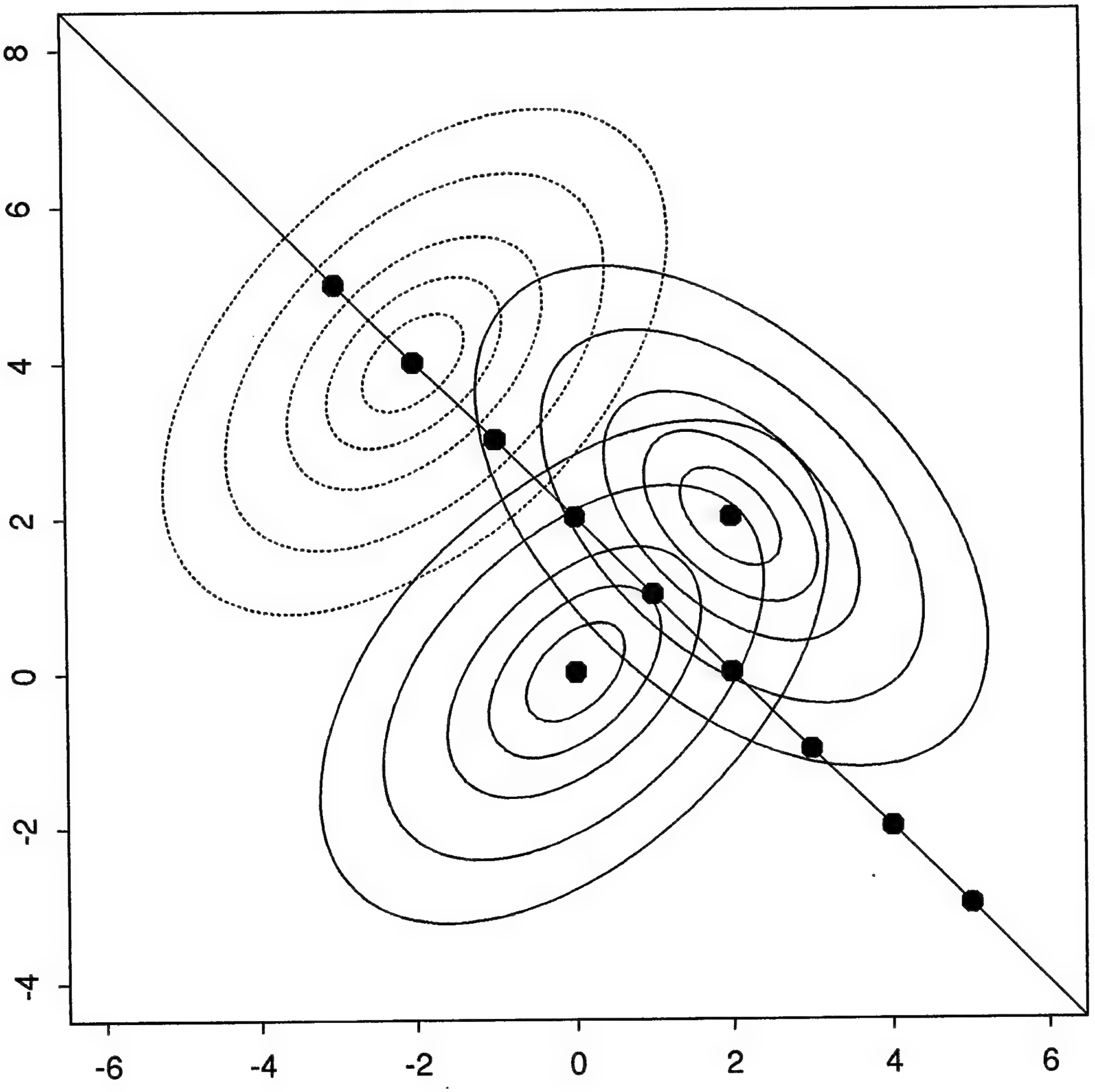
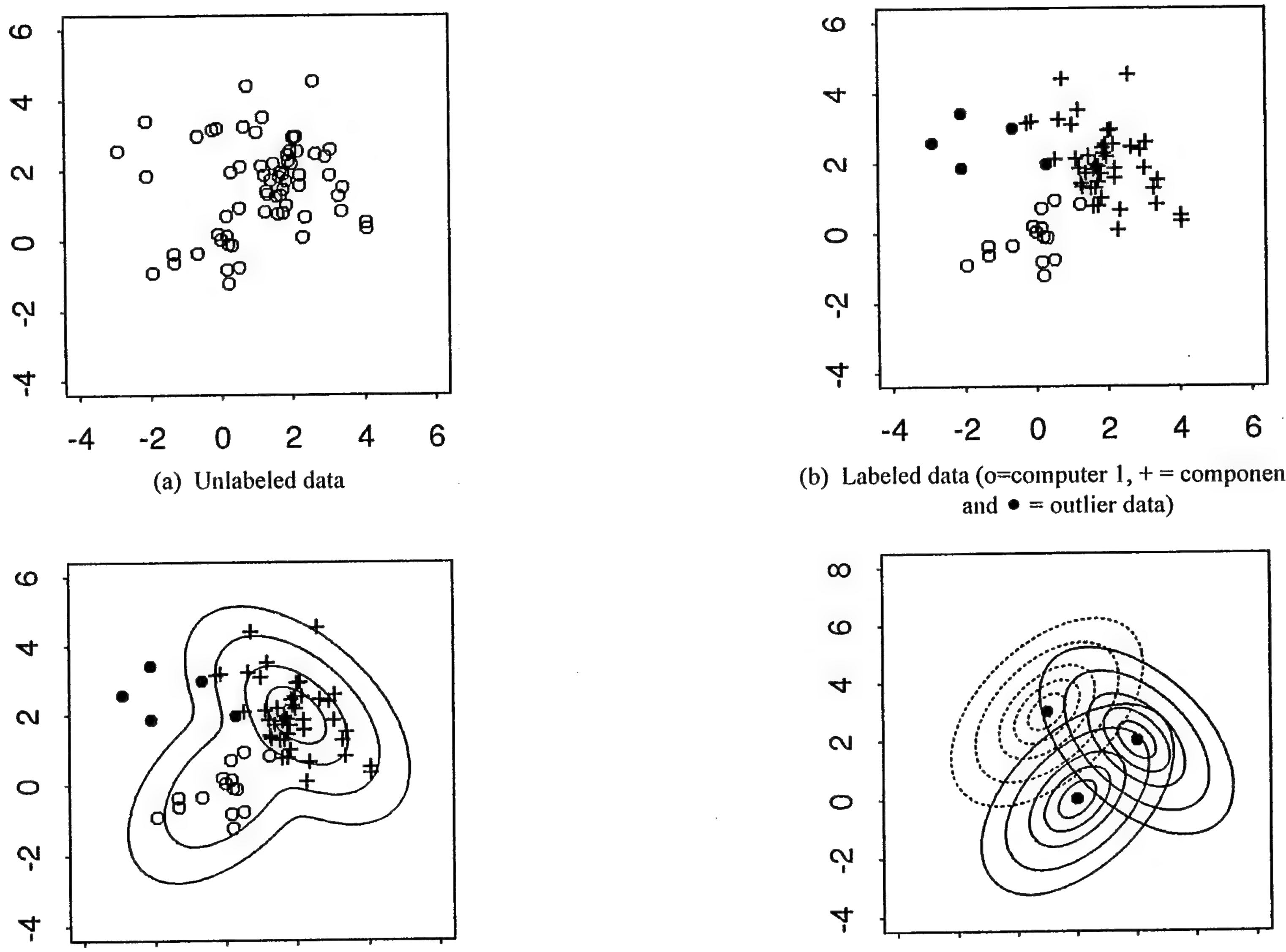


Figure 3 - Mixture distributions for Example $2b$ where $f(x) = .25N(\mu_1, \Sigma_1) + .75N(\mu_2, \Sigma_2)$



(a) Unlabeled data

(b) Labeled data (o=computer 1, + = component 2, and ● = outlier data)

(c) Labeled data long with contours of mixture population

(d) Contours of component populations and of outlier population (dotted)

Prof. Thomas Ahrens
Seismological Lab, 252-21
Division of Geological & Planetary Sciences
California Institute of Technology
Pasadena, CA 91125

Prof. Keiiti Aki
Center for Earth Sciences
University of Southern California
University Park
Los Angeles, CA 90089-0741

Prof. Shelton Alexander
Geosciences Department
403 Deike Building
The Pennsylvania State University
University Park, PA 16802

Dr. Thomas C. Bache, Jr.
Science Applications Int'l Corp.
10260 Campus Point Drive
San Diego, CA 92121 (2 copies)

Prof. Muawia Barazangi
Cornell University
Institute for the Study of the Continent
3126 SNEE Hall
Ithaca, NY 14853

Dr. Douglas R. Baumgardt
ENSCO, Inc
5400 Port Royal Road
Springfield, VA 22151-2388

Dr. T.J. Bennett
S-CUBED
A Division of Maxwell Laboratories
11800 Sunrise Valley Drive, Suite 1212
Reston, VA 22091

Dr. Robert Blandford
AFTAC/TT, Center for Seismic Studies
1300 North 17th Street
Suite 1450
Arlington, VA 22209-2308

Dr. Stephen Bratt
ARPA/NMRO
3701 North Fairfax Drive
Arlington, VA 22203-1714

Mr. Dale Breding
Sandia National Laboratories
Organization 9236, MS 0655
Albuquerque, NM 87185

Dr. Jerry Carter
Center for Seismic Studies
1300 North 17th Street
Suite 1450
Arlington, VA 22209-2308

Mr Robert Cockerham
Arms Control & Disarmament Agency
320 21st Street North West
Room 5741
Washington, DC 20451,

Dr. Zoltan Der
ENSCO, Inc.
5400 Port Royal Road
Springfield, VA 22151-2388

Dr. Stanley K. Dickinson
AFOSR/NM
110 Duncan Avenue
Suite B115
Bolling AFB, DC 20332-6448

Dr. Petr Firbas
Institute of Physics of the Earth
Masaryk University Brno
Jecna 29a
612 46 Brno, Czech Republic

Dr. Mark D. Fisk
Mission Research Corporation
735 State Street
P.O. Drawer 719
Santa Barbara, CA 93102

Dr. Cliff Frolich
Institute of Geophysics
8701 North Mopac
Austin, TX 78759

Dr. Holly Given
IGPP, A-025
Scripps Institute of Oceanography
University of California, San Diego
La Jolla, CA 92093

Dr. Jeffrey W. Given
SAIC
10260 Campus Point Drive
San Diego, CA 92121

Dan N. Hagedon
Pacific Northwest Laboratories
Battelle Boulevard
Richland, WA 99352

Dr. James Hannon
Lawrence Livermore National Laboratory
P.O. Box 808, L-205
Livermore, CA 94550

Dr. Roger Hansen
University of Colorado, JSPC
Campus Box 583
Boulder, CO 80309

Prof. David G. Harkrider
Phillips Laboratory
Earth Sciences Division, PL/GPE
29 Randolph Road
Hanscom AFB, MA 01731-3010

Prof. Danny Harvey
University of Colorado, JSPC
Campus Box 583
Boulder, CO 80309

Prof. Donald V. Helmberger
Division of Geological & Planetary Sciences
California Institute of Technology
Pasadena, CA 91125

Prof. Eugene Herrin
Geophysical Laboratory
Southern Methodist University
Dallas, TX 75275

Prof. Robert B. Herrmann
Department of Earth & Atmospheric Sciences
St. Louis University
St. Louis, MO 63156

Prof. Lane R. Johnson
Seismographic Station
University of California
Berkeley, CA 94720

Prof. Thomas H. Jordan
Department of Earth, Atmospheric &
Planetary Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139

Mr. Robert C. Kemerait
ENSCO, Inc.
445 Pineda Court
Melbourne, FL 32940

U.S. Dept of Energy
Max Koontz, NN-20, GA-033
Office of Research and Develop.
1000 Independence Avenue
Washington, DC 20585

Dr. Richard LaCoss
MIT Lincoln Laboratory, M-200B
P.O. Box 73
Lexington, MA 02173-0073

Prof. Charles A. Langston
Geosciences Department
403 Deike Building
The Pennsylvania State University
University Park, PA 16802

Jim Lawson, Chief Geophysicist
Oklahoma Geological Survey
Oklahoma Geophysical Observatory
P.O. Box 8
Leonard, OK 74043-0008

Prof. Thorne Lay
Institute of Tectonics
Earth Science Board
University of California, Santa Cruz
Santa Cruz, CA 95064

Dr. William Leith
U.S. Geological Survey
Mail Stop 928
Reston, VA 22092

Mr. James F. Lewkowicz
Phillips Laboratory/GPE
29 Randolph Road
Hanscom AFB, MA 01731-3010(2 copies)

Dr. Gary McCartor
Department of Physics
Southern Methodist University
Dallas, TX 75275

Prof. Thomas V. McEvilly
Seismographic Station
University of California
Berkeley, CA 94720

Dr. Keith L. McLaughlin
S-CUBED
A Division of Maxwell Laboratory
P.O. Box 1620
La Jolla, CA 92038-1620

Prof. Bernard Minster
IGPP, A-025
Scripps Institute of Oceanography
University of California, San Diego
La Jolla, CA 92093

Prof. Brian J. Mitchell
Department of Earth & Atmospheric Sciences
St. Louis University
St. Louis, MO 63156

Mr. Jack Murphy
S-CUBED
A Division of Maxwell Laboratory
11800 Sunrise Valley Drive, Suite 1212
Reston, VA 22091 (2 Copies)

Dr. Keith K. Nakanishi
Lawrence Livermore National Laboratory
L-025
P.O. Box 808
Livermore, CA 94550

Prof. John A. Orcutt
IGPP, A-025
Scripps Institute of Oceanography
University of California, San Diego
La Jolla, CA 92093

Dr. Howard Patton
Lawrence Livermore National Laboratory
L-025
P.O. Box 808
Livermore, CA 94550

Dr. Frank Pilotte
HQ AFTAC/TT
1030 South Highway A1A
Patrick AFB, FL 32925-3002

Dr. Jay J. Pulli
Radix Systems, Inc.
6 Taft Court
Rockville, MD 20850

Prof. Paul G. Richards
Lamont-Doherty Earth Observatory
of Columbia University
Palisades, NY 10964

Mr. Wilmer Rivers
Teledyne Geotech
1300 17th St N #1450
Arlington, VA 22209-3803

Dr. Alan S. Ryall, Jr.
Lawrence Livermore National Laboratory
P.O. Box 808, L-205
Livermore, CA 94550

Dr.Chandan K. Saikia
Woodward Clyde- Consultants
566 El Dorado Street
Pasadena, CA 91101

Mr. Dogan Seber
Cornell University
Inst. for the Study of the Continent
3130 SNEE Hall
Ithaca, NY 14853-1504

Secretary of the Air Force
(SAFRD)
Washington, DC 20330

Office of the Secretary of Defense
DDR&E
Washington, DC 20330

Thomas J. Sereno, Jr.
Science Application Int'l Corp.
10260 Campus Point Drive
San Diego, CA 92121

Dr. Michael Shore
Defense Nuclear Agency/SPSS
6801 Telegraph Road
Alexandria, VA 22310

Prof. David G. Simpson
IRIS, Inc.
1616 North Fort Myer Drive
Suite 1050
Arlington, VA 22209

Dr. Jeffrey Stevens
S-CUBED
A Division of Maxwell Laboratory
P.O. Box 1620
La Jolla, CA 92038-1620

Prof. Brian Stump
Los Alamos National Laboratory
EES-3
Mail Stop C-335
Los Alamos, NM 87545

Prof. Tuncay Taymaz
Istanbul Technical University
Dept. of Geophysical Engineering
Mining Faculty
Maslak-80626, Istanbul Turkey

Prof. M. Nafi Toksoz
Earth Resources Lab
Massachusetts Institute of Technology
42 Carleton Street
Cambridge, MA 02142

Dr. Larry Turnbull
CIA-OSWR/NED
Washington, DC 20505

Dr. Karl Veith
EG&G
2341 Jefferson Davis Highway
Suite 801
Arlington, VA 22202-3809

Prof. Terry C. Wallace
Department of Geosciences
Building #77
University of Arizona
Tuscon, AZ 85721

Dr. William Wortman
Mission Research Corporation
8560 Cinderbed Road
Suite 700
Newington, VA 22122

ARPA, OASB/Library
3701 North Fairfax Drive
Arlington, VA 22203-1714

HQ DNA
ATTN: Technical Library
Washington, DC 20305

Defense Technical Information Center
8725 John J. Kingman Road
Ft Belvoir, VA 22060-6218
(2 copies)

TACTEC
Battelle Memorial Institute
505 King Avenue
Columbus, OH 43201 (Final Report)

Phillips Laboratory
ATTN: GPE
29 Randolph Road
Hanscom AFB, MA 01731-3010

Phillips Laboratory
ATTN: TSML
5 Wright Street
Hanscom AFB, MA 01731-3004

Phillips Laboratory
ATTN: PL/SUL
3550 Aberdeen Ave SE
Kirtland, NM 87117-5776 (2 copies)

Dr. Michel Campillo
Observatoire de Grenoble
I.R.I.G.M.-B.P. 53
38041 Grenoble, FRANCE

Dr. Kin Yip Chun
Geophysics Division
Physics Department
University of Toronto
Ontario, CANADA

Prof. Hans-Peter Harjes
Institute for Geophysic
Ruhr University/Bochum
P.O. Box 102148
4630 Bochum 1, GERMANY

Prof. Eystein Husebye
IFJF
Jordskjelvstasjonen
Allegaten, 5007 BERGEN NORWAY

David Jepsen
Acting Head, Nuclear Monitoring Section
Bureau of Mineral Resources
Geology and Geophysics
G.P.O. Box 378, Canberra, AUSTRALIA

Ms. Eva Johannisson
Senior Research Officer
FOA
S-172 90 Sundbyberg, SWEDEN

Dr. Peter Marshall
Procurement Executive
Ministry of Defense
Blacknest, Brimpton
Reading FG7-FRS, UNITED KINGDOM

Dr. Bernard Massinon, Dr. Pierre Mechler
Societe Radiomana
27 rue Claude Bernard
75005 Paris, FRANCE (2 Copies)

Dr. Svein Mykkeltveit
NTNT/NORSAR
P.O. Box 51
N-2007 Kjeller, NORWAY (3 Copies)

Dr. Jorg Schlittenhardt
Federal Institute for Geosciences & Nat'l Res.
Postfach 510153
D-30631 Hannover , GERMANY

Dr. Johannes Schweitzer
Institute of Geophysics
Ruhr University/Bochum
P.O. Box 1102148
4360 Bochum 1, GERMANY

Trust & Verify
VERTIC
Carrara House
20 Embankment Place
London WC2N 6NN, ENGLAND